Interpreting disease genetics using functional genomics

Westra, Harm Jan

Interpreting disease genetics using functional genomics

Harm-Jan Westra

Interpreting disease genetics using functional genomics

Proefschrift

ter verkrijging van de graad
van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus
prof. dr. E. Sterken
en volgens besluit van
het College voor Promoties

De openbare verdediging zal
plaatsvinden op

woensdag 17 september 2014
om 12:45 uur

door

Harm-Jan Westra
Geboren op 23 april 1984
te Menaldumadeel

Promotores

Prof. dr. L. H. Franke
Prof. dr. C. Wijmenga

Beoordelingscommissie

Prof. dr. G. de Haan
Prof. dr. L.H. van den Berg
Prof. dr. R.K. Weersma

Contents

Introduction and outline

Adapted from: From genome to function by studying eQTLs,

BBA Biochimica et Biophysica Acta, 2014

1 From GWAS to SNP function

In the last few years, a large number of genome-wide association studies (GWASs) have been performed in attempts to uncover the genetic basis of many different complex diseases and traits. GWAS typically ascertain at least 300,000 common single nucleotide polymorphisms (SNPs) throughout the genome, and for each of these variants association with the disease is tested. For many traits, this approach has turned out to be highly successful; disease and trait associations for over 12,000 SNPs from over 1700 publications have now been reported (NHGRI Catalog of Published Genome-Wide Association studies)[1]. However, it soon became clear that the identified genetic variants typically explain only a very modest proportion of the total heritability of these traits.

One plausible explanation was that these GWAS had only investigated common SNPs (those with a minor allele frequency (MAF) above 5%). As such, many rare variants had not been ascertained, and it was therefore assumed that the common SNPs identified for a disease were actually tagging rarer variants (MAF < 5%) with a larger effect size. To test this hypothesis, fine-mapping studies were conducted, made possible with the availability of the next generation sequencing (NGS) methods: by sequencing candidate genes, whole exomes or genomes it is possible to identify rare variants[2] and their association with disease became testable through the development of dedicated oligonucleotide arrays that specifically target these rare variants (e.g. the ImmunoChip and MetaboChip). Although this helped to fine-map loci for various diseases, few rare variants have so far been identified that have a large effect size.

These results, along with the observation that many smaller-effect loci became genome-wide significant upon increasing the sample sizes used in many GWASs, suggested that the genetic architecture for many traits could well be highly polygenic. This was further supported by the availability of polygenic models in 2009[3,4]: these methods estimate the total proportion of variation that can be explained by all genotyped common SNPs, without requiring that any of the SNPs individually shows significant association (after correction for multiple testing). Initial results on adult height (which has an estimated heritability of 80% and is a phenotype that can be highly accurately quantified[4]) revealed that common genetic variants captured approximately 45% of the total variation in height, whereas the 180 genome-wide significant loci that had been found (when studying 180,000 samples) explained less than 10% of the variation in height. These results suggested that hundreds, or maybe even thousands, of genetic variants could well play a causal role in many traits.

These observations have proven highly problematic in trying to move from the discovery of these variants through GWAS to their biological interpretation for various reasons: given that many of the disease-causing variants are likely to be common, have small effect-sizes, and are often in near-perfect linkage disequilibrium (LD) with nearby SNPs. It is difficult to unequivocally identify the causal variant for each locus through traditional fine-mapping

1 Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9362–7 (2009).

2 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* 12, 745–55 (2011).

3 Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–52 (2009).

4 Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–9 (2010).

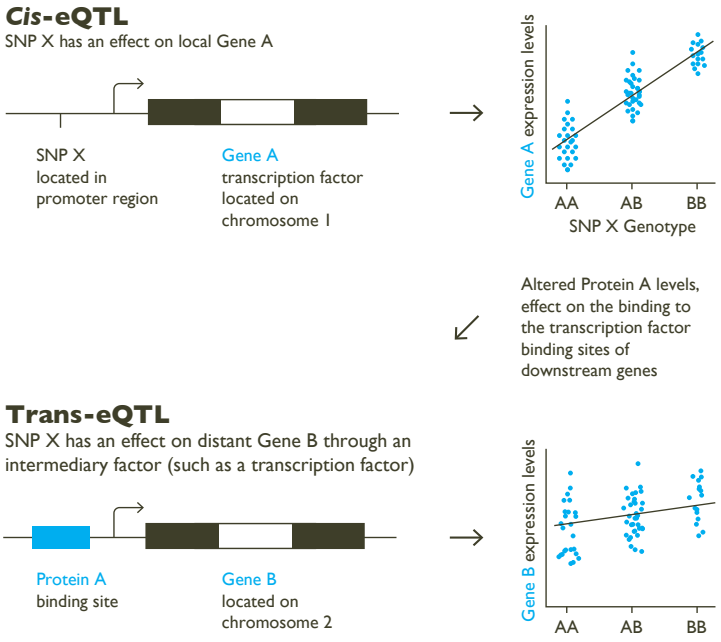


Figure 1.

eQTLs can be either local effects (*cis*-eQTLs), or distant, indirect effects (*trans*-eQTLs)

methods. This also strongly impairs the ability to accurately pinpoint the causal gene(s) in each locus. Additionally, the mechanisms and function of each of these trait-associated variants are largely unknown, since many of the trait-associated SNPs are not actually changing the protein structure (i.e. are non-synonymous or nonsense mutations), but are often located in non-coding regions of the genome. This suggests that these variants have a regulatory function. A compounding problem is that often tens of disease-associated variants have now been identified for many diseases, making it infeasible to knock-down, knock-out or over-express each of the genes within these loci.

In order to identify which genes are regulated by genetic variation, Jansen and Nap introduced the concept of 'genetical genomics'[5] in 2001: by correlating the genetic variants with intermediate molecular quantitative traits (such as gene expression levels, protein levels or methylation levels), it is possible to identify quantitative trait loci (QTLs). The first product of the genome, mRNA levels, can be quantified easily for thousands of genes at once, by either using microarrays or by conducting RNA-sequencing. It soon became clear that gene expression levels are strongly heritable: for all human genes the average heritability was estimated to be around 0.25[6–8]. Soon, expression QTL (eQTL) mapping was conducted in humans[9–11] (and model organisms such as *Arabidopsis Thaliana*[12], *Caenorhabditis Elegans*[13], mice and rats[14]), resulting in the identification of many genetic variants that affect gene expression levels.

5
Jansen, R. C. & Nap, J. P. Genetical genomics: the added value from segregation. *Trends Genet.* 17, 388–91 (2001).

6
Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* 39, 1217–24 (2007).

7
Monks, S.A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* 75, 1094–105 (2004).

8
Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* 39, 1202–7 (2007).

9
Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* 44, 1084–9 (2012).

10
Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* 7, e1002003 (2011).

11
Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* 43, 1193–201 (2011).

12
Keurentjes, J. J. B. *et al.* Regulatory network construction in Arabidopsis by using genome-widecgene expression quantitative trait loci. *Proc. Natl.Acad. Sci. U. S.A.* 104, 1708–13 (2007).

13
Viñuela, A., Snoek, L. B., Riksen, J.A. G. & Kammenga, J. E. Aging Uncouples Heritability and Expression-QTL in Caenorhabditis elegans. *G3 (Bethesda)*. 2, 597–605 (2012).

14
Tesson, B. M. & Jansen, R. C. eQTL analysis in mice and rats. *Methods Mol. Biol.* 573, 285–309 (2009).

15
Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57 (2009).

2 eQTLs as a means to functionally annotate trait associated SNPs

eQTLs can be divided into those that have local effects (*cis*-eQTLs), where the genetic variant is located near the affected gene (e.g. within 1 megabase), and those with distant effects (*trans*-eQTLs), where the genetic variant is located further away from the affected gene (e.g. more than 5 megabases away, or on a completely different chromosome; Figure 1).

2.1 *Cis*-eQTLs

Since *cis*-eQTLs often have a large effect size[15], relatively modest sample sizes permit the detection of *cis*-eQTLs for thousands of genes[6,16–20]. *Cis*-eQTL effects appear to be mostly additive effects[21], and *cis*-eQTL SNPs are often located close to the transcription start site (TSS) of genes or within gene bodies[22–24]. As the distance between the eQTL SNP and the TSS decreases, the eQTL effect size generally increases. *Cis*-eQTL SNPs that are located close to the TSS may alter transcription factor binding sites or other *cis*-regulatory elements (CREs), which in turn may affect transcription. The observation that *cis*-eQTL SNPs tend to be overlapping with activating CREs, such as DNAse-I hypersensitive sites (DHSs) and transcription factor binding sites, and tend to be depleted for repressive CREs (such as CTCF binding sites) strengthened this hypothesis[25]. Finally, trait-associated SNPs have been shown to be enriched for *cis*-eQTL effects[20,26–28] (Chapter 7 of this thesis), which further indicates that trait-associated SNPs

16 Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* 39, 1494–9 (2007).

17 Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* 7, e1002078 (2011).

18 Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848–53 (2007).

19 Fehrmann, R. S. N. *et al.* *Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* 7, e1002197 (2011).

20 Westra, H.-J. *et al.* Systematic identification of trans-eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–43 (2013).

21 Powell, J. E. *et al.* Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. *PLoS Genet.* 9, e1003502 (2013).

22 Stranger, B. E. & De Jager, P. L. Coordinating GWAS results with gene expression in a systems immunologic paradigm in autoimmunity. *Curr. Opin. Immunol.* 24, 544–551 (2012).

23 Veyrieras, J.-B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 4, e1000214 (2008).

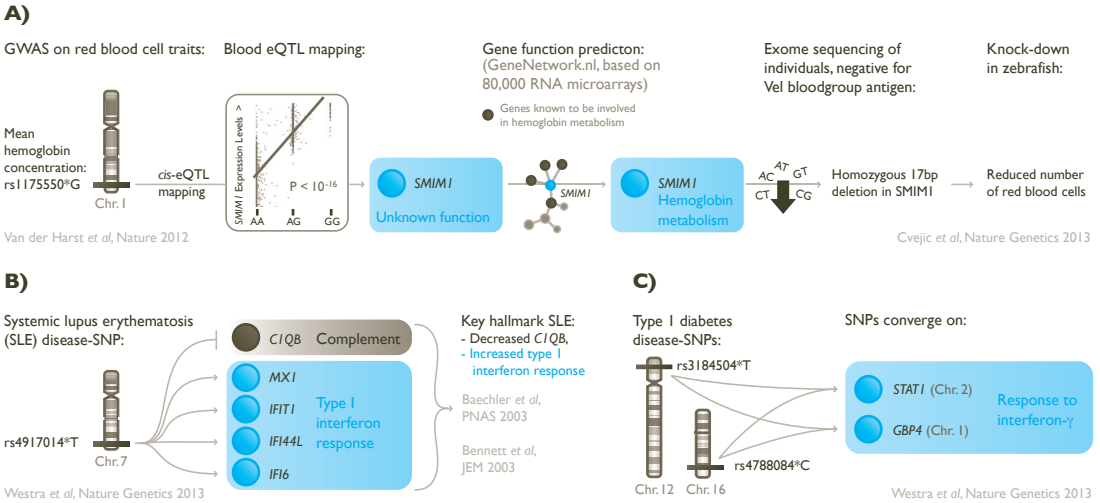


Figure 2.

Functional genomic studies translate GWAS findings into clear biological insight. A) A recent GWAS conducted on red blood cell traits identified a locus on chromosome 1 associated with mean hemoglobin concentration. Through subsequent *cis*-eQTL mapping and gene function prediction (using a compendium of 80,000 microarrays), *SMIM1* was identified as the possibly causal gene in the locus on chromosome 1 that was predicted to be involved in hemoglobin metabolism. Subsequent exome-sequencing revealed this gene underlies the rare Vel blood group, and knock-down of Vel1 in zebrafish resulted in a reduced number of red blood cells. B) Through *trans*-eQTL mapping in healthy individuals the downstream effects for the systemic lupus erythematosis (SLE) SNP rs4917014 were identified. These effects are identical to the key hallmarks of SLE: decreased complement 1q levels and an increased type 1 interferon response. C) SNPs that increase risk for the same disease 'converge' on the same downstream genes: two unlinked type 1 diabetes SNPs affect exactly the same downstream genes in trans (*STAT1* and *GBP4*, both involved in the interferon-γ response).

are often regulatory. *Cis*-eQTLs can aid in pinpointing the causal variant within a locus: after a GWAS on red blood cell traits[29], *cis*-eQTL mapping was performed in whole blood samples, which identified a *cis*-eQTL in the *SMIM1* locus on chromosome 1. Subsequent functional annotation using a gene expression co-regulation network suggested *SMIM1* was the causal gene within the locus. A follow-up exome sequencing study and knock down experiment in zebrafish revealed that *SMIM1* underlies the Vel blood group (Figure 2A).

Although *cis*-eQTLs, such as the *SMIM1* example, can provide valuable information about the likely causal gene for trait-associated SNPs, finding the causal gene underneath GWAS peaks is not always straightforward: LD might be so extensive that many candidate genes remain, or the regulatory regions that are influenced by the genetic variants may actually be located megabases away from the transcription start site of the causal gene. This has recently been observed for intronic variants within the *FTO* gene that have been found to be associated with Type 2 diabetes and obesity[31,32]. Surprisingly, these variants do not show a *cis*-eQTL effect on *FTO*, but they do affect the gene expression levels of *IRX3*, which is located megabases away from the *FTO* locus[33]. Knocking-out *IRX3* in mice results in a 30% weight decrease in mice, confirming the importance of *IRX3* in regulating weight. These results illustrate that the genes that are located in very close proximity to the associated variant are not always the causal gene and also that variants associated with GWAS may have functional consequences on genes located megabases away, which raises the question whether such effects should be considered *trans*-eQTLs.

2.2 *Trans*-eQTLs

In contrast to *cis*-eQTL effects, the effect sizes of *trans*-eQTLs are generally small[9,34]. As a consequence, the sample sizes required to detect such effects are large, and as a result, the number of reported *trans*-eQTLs has remained small[9,17,19,35–37] in comparison to the number of reported *cis*-eQTLs. However, initial *trans*-eQTL studies have shown that *trans*-eQTL analysis provides valuable insight into disease pathogenesis. For example, multiple *trans*-eQTL genes were previously identified that are affected by a single SNP that is associated with type 2 diabetes and high-density lipoprotein levels. SNPs associated with these *trans*-genes also showed genetic association with various metabolic phenotypes[37], indicating that *trans*-eQTL mapping is able to identify coherent networks of genes that are likely to be causally involved in disease pathogenesis. Similarly, *trans*-eQTL genes were identified that are affected by a SNP in the *IRF7* locus, associated with the auto-immune disease type 1 diabetes. These downstream *trans*-genes showed an association with viral response[36]. To detect more *trans*-eQTL effects, sample-sizes were increased by performing meta-analyses[19,20]: a meta-analysis of 1469 whole blood samples showed that HLA SNPs were 10-fold enriched for showing *trans*-eQTL effects. For a few different complex traits it was also shown that SNPs independently associated with these traits affected the expression of exactly the same downstream genes in trans, creating functional converging pathways that are relevant for the traits associated with these SNPs[19] (Chapter 3 of this thesis). A larger meta-analysis involving 5311 whole blood samples further

24 Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246–50 (2009).

25 Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative Modeling of eQTLs and *Cis*-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS Genet.* 9, e1003649 (2013).

26 Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6, e1000888 (2010).

27 Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6, e1000895 (2010).

28 Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302 (2010).

29 Van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* 492, 369–75 (2012).

30 Cvejic, A. *et al.* SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* 45, 542–5 (2013).

31 Frayling, T. M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316, 889–94 (2007).

32 Dina, C. *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat. Genet.* 39, 724–6 (2007).

33 Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371–375 (2014).

34 Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184–94 (2009).

35 Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44, 502–10 (2012).

36 Heinig, M. *et al.* A *trans*-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* 467, 460–4 (2010).

increased the number of reported *trans*-eQTL genes to 430 and showed that *trans*-eQTLs can be informative of disease pathogenesis: in two previous cross-sectional studies, several interferon response genes had been identified that show strongly dysregulated expression in the blood of systemic lupus erythematosus (SLE) patients (Figure 2B). The *trans*-eQTL study identified a single SNP, associated with SLE, that affected exactly these genes, indicating that dysregulation of these interferon response genes is already detectable when a healthy individual is carrying SLE susceptibility alleles[20] (Chapter 7 of this thesis). Similar to the meta-analysis in 1469 individuals, this larger meta-analysis provided information on the convergence of functional pathways, including converging effects originating from two type 1 diabetes associated variants, affecting the well-known type 1 diabetes gene *STAT1* (Figure 2C).

2.3 Cell type and tissue specificity

Gene expression levels often vary considerably between different tissues and cell types[38]. As such, eQTL mapping studies have now been performed in various cell-types and tissues, such as fibroblasts, liver[17,39–41], lung[42], brain[16], muscle[41], adipose tissue[41], skin[43], various purified blood cell types (e.g. lymphoblastoid cell-lines (LCLs)[10,40,43–45], B-cells[35], monocytes[35], and T-cells[24]), and whole blood[19,20,46]. Early comparisons between cell types showed that the number of shared eQTL effects varies widely with the cell types or tissues under study. A comparison of skin and LCL eQTLs showed that 70% of *cis*-eQTLs found in LCLs could also be detected in skin[43], while a comparison of fibroblasts, T-cells and B-cells showed an overlap of up to 12% of the detected *cis*-eQTLs in any combination of two of these three cell types[24]. However, these studies had overestimated the cell-type specificity because of their small sample size and the statistical methods employed to make these comparisons. A more recent comparison of B-cells and monocytes in over 280 individuals showed a higher overlap: 21.8% of the detected *cis*-eQTLs and 7% of the detected *trans*-eQTLs were shared between both cell types[35], which suggests that genetic regulation in trans is more cell-type-specific than *cis* regulation. Only a small proportion of the identified eQTL effects in this study could be replicated in whole blood (even though blood is partly comprised of monocytes and B-cells), indicating that eQTL mapping in a tissue that is composed of many cell types may reduce the power to find cell-type-specific eQTL effects.

Another recent study, comparing five tissues (subcutaneous and visceral adipose tissue, muscle, liver, and whole blood), described how 28.7% of the *cis*-eQTLs show differences across tissues[41]. Of these, 33% had eQTL effects unique to one of the tissues, 47.9% showed eQTLs originating from different SNPs in different tissues, and 4.4% unexpectedly showed a different direction of effect in one or more tissues, something that has recently been observed in other studies as well[35,47]. This study also showed *cis*-eQTL effects for 46.4% of the tested trait-associated SNPs, and indicated that these SNPs are enriched for tissue-dependent effects, compared to frequency matched SNPs. Another study on the same tissue samples showed that this tissue specificity is not limited to protein coding genes, but is also present for long intergenic non-coding RNAs[48] (Chapter 5 of this thesis). eQTLs that are shared across tissues and cell types have larger

37 Small, K. S. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat. Genet.* 43, 561–4 (2011).

38 Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4465–70 (2002).

39 Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6, e107 (2008).

40 Zhong, H. *et al.* Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* 6, e1000932 (2010).

41 Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* 8, e1002431 (2012).

42 Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* 8, e1003029 (2012).

43 Ding, J. *et al.* Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in *cis*-eQTL signals. *Am. J. Hum. Genet.* 87, 779–89 (2010).

44 Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* 8, e1002639 (2012).

45 Choy, E. *et al.* Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* 4, e1000287 (2008).

46 Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* 21, 48–54 (2013).

47 Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. *Nature* 505, 208–11 (2014).

48 Kumar, V. *et al.* Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. *PLoS Genet.* 9, e1003201 (2013).

effect sizes, and their associated SNPs are located closer to the TSS[25,41] than tissue- and cell-type specific eQTLs. On average, 29% of the *cis*-eQTL loci also appear to have multiple independent SNPs affecting the same transcript[25]. Overall, these studies show that the genetic regulation of gene expression is complex and differs across cell-types and tissues, especially for disease-associated genetic variants.

2.4 Context specificity

Apart from differences in cell types, a large fraction of gene expression variation is due to the effect of environmental factors, begging the question whether some of these environmental factors might induce eQTLs. Several environmental factors have now been assessed in the context of eQTLs, which include response to radiation[49], geography[50], different treatments for disease[51,52], response to influenza vaccination[53], and infections with tuberculosis[54] and malaria[55]. However, the sample sizes for these studies have generally been rather small (up to 194 individuals), due to the difficulties and costs involved in acquiring samples that are relevant for the specific environmental factor. More powerful studies have been published as well: a study in monocytes from 1490 independent individuals showed 651 *cis*-eQTLs that have interactions with either age, smoking status, gender, blood pressure and lipid traits[56]. However, the early stage of the studies of context-specific eQTLs has sometimes led to discrepant results being observed: for example, a gender-stratified analysis in 379 LCLs suggested that between 12% and 15% of the autosomal eQTLs function in a sex-biased manner[57], and a larger study in peripheral blood samples from 1240 individuals showed interactions for 623 eQTLs with age[58]. Although the large fraction of gender mediated effects in the LCLs may be caused by (epi-)genomic alterations caused by the Eppstein Barr virus immortalization of these cell-lines[59], a subsequent, but much larger study of 5254 peripheral blood samples showed only 14 and 10 eQTLs interacting with gender and age, respectively[60]. One potential explanation for these discrepancies could be the statistical challenges, associated with performing large-scale gene environment interaction analysis. In order to get robust significance estimates of interaction effects, heteroscedasticity-consistent standard errors should be used[61] (e.g. available through the R package 'Sandwhich').

Still, context-specific eQTL studies hold great promise. For instance, it should be possible to use whole blood eQTL data to make inferences about the specific cell-types in which these eQTLs manifest themselves, by using the abundance of such cell types as an interaction term (Chapter 3 of this thesis). Additionally, a recent study in monocytes, comparing the effect sizes of eQTLs before and after stimulation with interferon-γ and bacterial lipopolysaccharides (LPS; which was measured at two different time-points), reported that 51.4% of the eQTLs detected before stimulation were not detectable after stimulation, sometimes in a time-dependent manner[62]. Additionally, a study assessing the effect of the stimulation of dendritic cells with LPS, influenza and interferon-β, showed 121 eQTLs associated with changes in gene expression due to these stimuli (response-eQTLs; *cis*-reQTLs)[63]. Like cell-type specific effects, stimulus dependent eQTLs appeared to have a larger distance between the SNP and

49 Smirnov, D.A., Morley, M., Shin, E., Spielman, R. S. & Cheung, V. G. Genetic analysis of radiation-induced changes in human gene expression. *Nature* 459, 587–91 (2009).

50 Idaghdour, Y. *et al.* Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat. Genet.* 42, 62–7 (2010).

51 Romanoski, C. E. *et al.* Systems genetics analysis of gene-by-environment interactions in human cells. *Am. J. Hum. Genet.* 86, 399–410 (2010).

52 Maranville, J. C. *et al.* Interactions between glucocorticoid treatment and *cis*-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genet.* 7, e1002162 (2011).

53 Franco, L. M. *et al.* Integrative genomic analysis of the human immune response to influenza vaccination. *Elife* 2, e00299 (2013).

54 Barreiro, L. B. *et al.* Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.* 109, 1204–9 (2012).

55 Idaghdour, Y. *et al.* Evidence for additive and interaction effects of host genotype and infection in malaria. *Proc. Natl. Acad. Sci. U. S. A.* 109, 16786–93 (2012).

56 Zeller, T. *et al.* Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One* 5, e10693 (2010).

57 Dimas, A. S. *et al.* Sex-biased genetic effects on gene regulation in humans. *Genome Res.* 22, 2368–75 (2012).

58 Kent, J. W. *et al.* Genotype×age interaction in human transcriptional ageing. *Mech. Ageing Dev.* 133, 581–90 (2012).

59 Hansen, K. D. *et al.* Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Res.* 24, 177–84 (2014).

60 Yao, C. *et al.* Sex- and age-interacting eQTLs in human complex diseases. *Hum. Mol. Genet.* ddt582– (2013). doi:10.1093/hmg/ddt582

61 White, H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48, 817–838 (1980).

the transcript compared to effects shared with unstimulated cells[62], and can affect specific transcription factor binding sites[63]. Both studies showed that trait-associated SNPs can have stimulus dependent effects, which provides further insight in the downstream effects of disease associated SNPs[62,63].

2.5 RNA-sequencing

So far, most eQTL mapping studies have measured gene expression levels using microarray technology. With the advent of NGS, the sequencing of RNA molecules (RNA-seq) has also become feasible. RNA-seq has a much larger dynamic range than microarray based gene expression quantification, and as such, a smaller amount of RNA molecules is required to accurately quantify gene expression levels[64,65]. The initial eQTL mapping studies performed using RNA sequencing data on LCLs have shown that the gene expression measurements between microarrays and RNA sequencing data generally correlated well (with correlations ranging between 0.6 and 0.781)[66]. As such, *cis*-eQTLs detected using RNA-seq replicated well when using microarray data, with up to 70% of the *cis*-eQTLs detected on microarrays being replicated in the Nigerian HapMap population[66], and up to 81% being replicated in a Central European HapMap population[67]. RNA-seq allows for a higher resolution of gene expression quantification than microarrays, since RNA-seq is not limited to a predefined set of oligonucleotide probes. Consequently, the RNA-seq studies on LCLs showed that *cis*-eQTL effects are not limited to annotated genes: in the Nigerian HapMap population, for example, the expression of 4031 unannotated transcripts was reported[66]. The higher resolution of RNA-seq also allows for better estimation of the correlation structure between exons and can thus be used to impute missing gene expression data for exons or transcripts[67], and it allows for better mapping of *cis*-eQTLs within exons. Comparing RNA-seq with an earlier microarray-based study on the same samples, RNA-seq-based eQTL mapping was better able to detect exon *cis*-eQTLs, most of which were located in genes with a high level of transcription, which indicates that RNA-seq is less prone to saturation of the gene expression signal, and that splicing complexity is not properly picked up by microarray-based studies[67]. Apart from exon-based *cis*-eQTLs, the relative ratios of different transcript isoforms can also be used as a quantitative trait in RNA-seq-based studies, in order to detect splicing-QTLs (sQTLs): 187 and 110 significant sQTLs were detected in the Nigerian and Central European HapMap populations, respecively[66,67], 639 genes were detected with significant sQTLs in a more recent LCL based study of 462 individuals[68], and 2851 sQTLs were detected in a whole blood study of 922 individuals[69], which indicates SNPs also regulate gene expression through altering different transcript isoforms. sQTLs appear to originate from different regulatory variants than eQTLs, since sQTL SNPs show less enrichment near the 5' end of genes compared to *cis*-eQTLs[69], but more enrichment in splice sites, 3' untranslated regions (3' UTR) and promotors[68]. Additionally, 57% of the eQTL genes that also showed an sQTL had an independent effect when conditioning for sQTLs, further indicating the independence between eQTL and sQTL regulation[68]. Overall, these studies show that genetic variation has a smaller influence on splicing than on overall gene expression. Finally, different RNA sequencing strategies can be used to answer

62 Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* 343, 1246949 (2014).

63 Lee, M. N. H. M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* (80-.). 343, 1246980–1246980 (2014).

64 Sun, W. & Hu, Y. eQTL Mapping Using RNA-seq Data. *Stat. Biosci.* 5, 198–219 (2013).

65 't Hoen, P. A. C. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* 31, 1015–22 (2013).

66 Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–72 (2010).

67 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–7 (2010).

68 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–11 (2013).

69 Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24, 14–24 (2014).

70 Zhernakova, D.V *et al.* DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* 9, e1003594 (2013).

different biological questions. For example, DeepSAGE, a sequencing strategy that uses primer sequences that specifically target the 3' ends of genes, is more suitable for detecting gene expression variation near the 3' ends than conventional RNA-seq, which, due to its random hexamer library design, shows larger fragmentation near the ends of genes[70]. A study applying the DeepSAGE method showed 12 poly-adenylation QTLs, that transcripts more often have an altered 3'UTR length, but also showed that different RNA-seq variants can be successfully meta-analyzed[70] (Chapter 6 of this thesis).

Outline of this thesis

This thesis has two main aims: 1) to present novel computational methods that can be used to improve the statistical power to detect genetic effects on gene expression, and 2) to assess the effects of genotypic variants on gene expression, in order to determine the downstream effects of trait-associated and other genetic variants.

Part 1 – Computational methods for eQTL mapping

Chapter 2 describes the *MixupMapper* program, which is a method to detect sample mix-ups in genetical genomics datasets. We applied this method to six published datasets that had both genotype and gene expression data, and observed that, on average, 3% of all samples in these published datasets had been mixed up. We modeled the impact of these mix-ups on the power to detect genetic effects, and showed that correcting such mix-ups increases the number of detectable eQTLs.

Chapter 3 describes additional methods to improve the statistical power to detect eQTLs. We used principal component analysis to correct the gene expression data for technical artifacts and combined two datasets in an eQTL meta-analysis of 1,469 whole blood samples. We observed an increase of the number of detected *trans*-eQTLs, and showed that independent trait-associated SNPs can converge on the same genes in *trans*.

Chapter 4 describes a method to predict the cell type specificity of eQTLs that have been detected in a tissue that consists of many different cell types, such as whole blood. The method uses known cell type proportions in a training dataset to create proxy phenotypes, and uses an interaction model to predict the interaction of the genetic effect with the proxy phenotype. Using this approach, we created a proxy for neutrophils, which we applied to 5,683 whole blood samples. Our method predicted that approximately 7% of the *cis*-eQTLs in whole blood are specific for neutrophils, and validated this hypothesis in six cell-type-specific eQTL datasets, and showed that SNPs associated with Crohn's disease preferentially affect gene expression levels within neutrophils.

Part 2 – eQTL association studies

Chapter 5 describes a study of the genetic effects on the expression of large intergenic non-coding RNAs (lincRNA) in five different tissues. We showed that variants often alter lincRNA gene expression levels and that lincRNA eQTLs are often tissue-dependent.

Chapter 6 describes an eQTL study using DeepSAGE RNA-sequencing, which showed that genetic variants can alter the length of the poly-adenylation of transcripts. This study also showed that the RNA-seq technique is more powerful in detecting eQTL effects than microarrays and that different RNA-seq variants can be effectively meta-analyzed to further increase the statistical power.

Chapter 7 describes a large *cis*- and *trans*-eQTL meta-analysis of 5,311 whole blood samples from seven independent cohorts. This study showed that *trans*-eQTLs can be detected and replicated in independent cohorts and showed that trait-associated SNPs are enriched for both *cis*- and *trans*-eQTLs. This study also identified additional convergent effects for trait-associated SNPs and indicated that independent trait-associated SNPs that affect the same transcription factor in *cis* can have different effects in *trans*.

Part 3 – Discussion

Chapter 8 places the work described in this thesis in the broader perspective of genetical genomics studies. We describe the implications of the findings of this thesis with respect to the genetics of complex traits and suggest future perspectives for the analysis of the genetics of gene expression as well as for functional genetics studies in general.

Introduction and outline

Part

MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects

Bioinformatics, 2011 May; 27(15): 2104-2111

Harm-Jan Westra[1], Ritsert C. Jansen[2], Rudolf S. N. Fehrmann[1], Gerard J. te Meerman[1], David van Heel[3,†], Cisca Wijmenga[1,†] and Lude Franke[1,3,†,*]

2

1 Department of Genetics, University Medical Center Groningen
2 Groningen Bioinformatics Center, Groningen
Biomolecular Sciences and Biotechnology Institute, University of Groningen, 9700AB, Groningen The Netherlands
3 Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK

Abstract

Motivation

Sample mix-ups can arise during sample collection, handling, genotyping or data management. It is unclear how often sample mix-ups occur in genome-wide studies, as there currently are no post-hoc methods that can identify these mix-ups in unrelated samples. We have therefore developed an algorithm (*MixupMapper*) that can both detect and correct sample mix-ups in genome-wide studies that study gene expression levels.

Results

We applied *MixupMapper* to five publicly available human genetical genomics datasets. On average, 3% of all analyzed samples had been assigned incorrect expression phenotypes: in one of the datasets 23% of the samples had incorrect expression phenotypes. The consequences of sample mix-ups are substantial: when we corrected these sample mix-ups, we identified on average 15% more significant *cis*-expression quantitative trait loci (*cis*-eQTLs). In one dataset, we identified three times as many significant *cis*-eQTLs after correction. Furthermore, we show through simulations that sample mix-ups can lead to an underestimation of the explained heritability of complex traits in genome-wide association datasets.

Availability and implementation

MixupMapper is freely available at http://www.genenetwork.nl/mixupmapper/

Contact

lude@ludesign.nl

Supplementary Information

Supplementary information is available online at http://www.genenetwork.nl/mixupmapper/

1 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9362–7 (2009).

2 Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302 (2010).

3 Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* 41, 703–7 (2009).

4 Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* 42, 579–89 (2010).

5 Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42, 1118–25 (2010).

6 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–13 (2010).

7 Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42, 937–48 (2010).

8 Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–8 (2010).

9 Jansen, R. C. & Nap, J. P. Genetical genomics: the added value from segregation. *Trends Genet.* 17, 388–91 (2001).

10 Buyske, S., Yang, G., Matise, T. C. & Gordon, D. When a case is not a case: effects of phenotype misclassification on power and sample size requirements for the transmission disequilibrium test with affected child trios. *Hum. Hered.* 67, 287–92 (2009).

Introduction

Genome-wide studies have identified many disease-associated variants for a wide plethora of complex human diseases[1] (such as celiac disease[2], type 1 diabetes[3] and type 2 diabetes[4], Crohn's disease[5]), and complex continuous phenotypes (such as lipid levels[6], body mass index (BMI)[7] and height[8]). Many of these studies[2,4,6–8] also assess the effect of the identified genetic variants on gene expression variation (i.e. genetical genomics[9]), by mapping expression quantitative trait loci (eQTL). As such, these studies involve many steps before actual analysis of the data, during each of which samples could be accidentally swapped. Since these studies are pushing towards larger sample-sizes in order to be able to identify ever smaller effects, the presence of sample mix-ups becomes almost unavoidable.

It is known from simulations that sample mix-ups can have an effect on the power to detect genetic associations in genome-wide studies[10–14], which may present a problem to detect variants with small effects. However, it is unclear how often such sample mix-ups actually occur in studies investigating gene expression.

The common method to detect sample mix-ups in genome-wide association studies (GWAS) is to check for heterozygous genotypes for X-chromosomal markers in males. However, this procedure will not identify sample mix-ups between samples of identical gender. While it is also possible to use multiple phenotypes that can be well predicted based on genetic markers (such as eye color[15], hair color[15] and ABO blood group[16]), we are not aware of any study where this has been applied to identify sample mix-ups in GWAS. It is obvious that if there would be a considerable number of such phenotypes available, identification of nearly all sample mix-ups should become feasible. Another method to prevent sample mix-ups that has is commonly used in GWA studies involves the genotyping of a small number of variants prior to hybridization to the chip. Post-hoc concordance analysis then allows to resolve mixed-up samples, although this method does not resolve mix-ups that might have been introduced during phenotyping. Although these methods are tailored for GWAS, they are also applicable to the genetical genomics datasets. This however does not apply to the gene expression data, for which to our knowledge no methods to detect sample mix-ups currently exist.

Our method (*MixupMapper*) uses gene expression levels for genes which are influenced by genetic variation located near these genes (*cis*-eQTLs). On the basis of such *cis*-eQTL effects, our method measures the difference between actual gene expression levels and predicted expression levels that are solely based on genotype data of *cis*-SNPs. Using this distance measure, MixupMapper is able to detect and correct sample mix-ups with high sensitivity and specificity. In this study, we analyzed five publicly available human genome-wide studies on gene expression and observe that sample mix-ups are frequent. Subsequently,

11 Gordon, D. & Finch, S. J. Consequences of error. Encycl. Genet. genomics, proteomics *Bioinforma.* (2006).

12 Ho, L. A. & Lange, E. M. Using public control genotype data to increase power and decrease cost of case-control genetic association studies. *Hum. Genet.* 128, 597–608 (2010).

13 Samuels, D. C., Burn, D. J. & Chinnery, P. F. Detecting new neurodegenerative disease genes: does phenotype accuracy limit the horizon? *Trends Genet.* 25, 486–8 (2009).

14 Zheng, G. & Tian, X. The impact of diagnostic error on testing genetic association in case-control studies. Stat. Med. 24, 869–82 (2005).

15 Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* 39, 1443–52 (2007).

16 YIP, S. P. Sequence variation at the human ABO locus. *Ann. Hum. Genet.* 66, 1–27 (2002).

17 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* 461, 747–53 (2009).

18 Choy, E. *et al.* Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* 4, e1000287 (2008).

19 Heinzen, E. L. *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.* 6, e1 (2008).

20 Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848–53 (2007).

21 Webster, J. A. *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.* 84, 445–58 (2009).

Table 1.

Analyzed genetical genomics datasets

Trait	Population	Sample size	Platform	Accession
Choy *et al*	HapMap CHB+JPT	87	Affymetrix HG-U133A	GSE11582
	HapMap CEU	84	Affymetrix HG-U133A	
	HapMap YRI	85	Affymetrix HG-U133A	
Stranger *et al*	HapMap CHB+JPT	90	Illumina Sentrix Human6 Beadchip	GSE6536
	HapMap CEU	90	Illumina Sentrix Human6 Beadchip	
	HapMap YRI	90	Illumina Sentrix Human6 Beadchip	
Zhang *et al*	HapMap CEU	87	Affymetrix Human ST1.0 Exon array	GSE9703
	HapMap YRI	89	Affymetrix Human ST1.0 Exon array	
Webster *et al*	Brain	363	Illumina Human Refseq-8	GSE15222
Heinzen *et al*	Brain	93	Affymetrix Human ST1.0 Exon array	†
	PBMC	80	Affymetrix Human ST1.0 Exon array	
Wolfs *et al*	Liver	73	Illumina Human HT12v3	GSE22070 *

† Data available through: http://people.genome.duke.edu/~dg48/SNPExpress. * We excluded the Wolfs,M.G. *et al.* (unpublished data) dataset from eQTL mapping, since this dataset was not previously published as a genetical genomics dataset.

we show that correcting sample mix-ups can yield a substantial increase in the number of significant *cis*-eQTLs. Furthermore, we show through simulations that sample mix-ups have large effects in genome-wide association studies as well, when detecting genome-wide significant associations, which may account for a considerable proportion of the missing heritability problem which affects many current GWAS studies[17].

Methods

Datasets

We used five genetical genomics studies[18–22] to assess the prevalence of sample mix-ups (Table 1). To our knowledge, this list includes all publicly available datasets that include both genome-wide genotype and gene expression data (as of October 2010). For the various HapMap datasets[18,20,22] we confined ourselves to the 309,565 SNPs present on the commonly used Illumina HumanHap300 platform (to limit the number of calculations). The studies that investigated samples from the HapMap project concentrated on the Central European (CEU), Chinese (CHB), Japanese (JPT) and Yoruban (YRI) populations. We combined the CHB and JPT populations since their sample sizes were very small and both reflect Asian samples. As such, we analyzed three sample sets for the studies that used HapMap samples (CEU, CHB+JPT and YRI) for the datasets of Stranger et al and Choy *et al.* We analyzed two sets of samples for Zhang et al's dataset as they had only investigated the CEU and YRI subpopulations. The dataset from Heinzen et al consisted of two separate sets of samples from peripheral blood mononuclear cells (PBMCs) and brain tissue that were analyzed separately. Finally, we included a dataset on brain tissue samples from Webster *et al.* We also assessed a liver dataset from Wolfs et al, but did not include eQTL mapping results, as this dataset was not published as a genetical genomics dataset before.

Cis-eQTL mapping

For the sample mix-up analysis, we ran an initial *cis*-eQTL analysis on each dataset. Although we expected the presence of sample mix-ups to have a large effect on the ability to detect *cis*-eQTLs, we assumed this influence was limited for the *cis*-eQTLs with the strongest effects. Gene expression datasets were quantile normalized and $\log_2$ transformed, if appropriate, prior to eQTL mapping. *Cis*-eQTL mapping was performed by using Spearman rank correlations (minor allele frequency (MAF) > 5%, a Hardy-Weinberg equilibrium (HWE) P-value ≥ 0.0001, SNP call-rate ≥ 95%). Only those SNP-probe pairs were tested that were within a vicinity of 250 kilobases (kb). Multiple testing correction was performed by controlling the false discovery rate (FDR) at 0.05 by permuting the phenotype to genotype sample labels (swapping sample phenotype labels, thus preserving the correlation structure within both the genotype and expression data[23]) and re-running the eQTL mapping 1,000 times. The numbers of *cis*-eQTLs that we have reported here refer to the numbers of unique probes that show a *cis*-association. It is important to note that we did not correct for any potentially false-positive *cis*-eQTLs caused by primer polymorphisms within the expression probe because they actually assist in determining the correct correspondence between genotype and gene expression

22 Zhang, W. *et al.* Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum. Genet.* 125, 81–93 (2009).
23 Breitling, R. *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 4, e1000232 (2008).

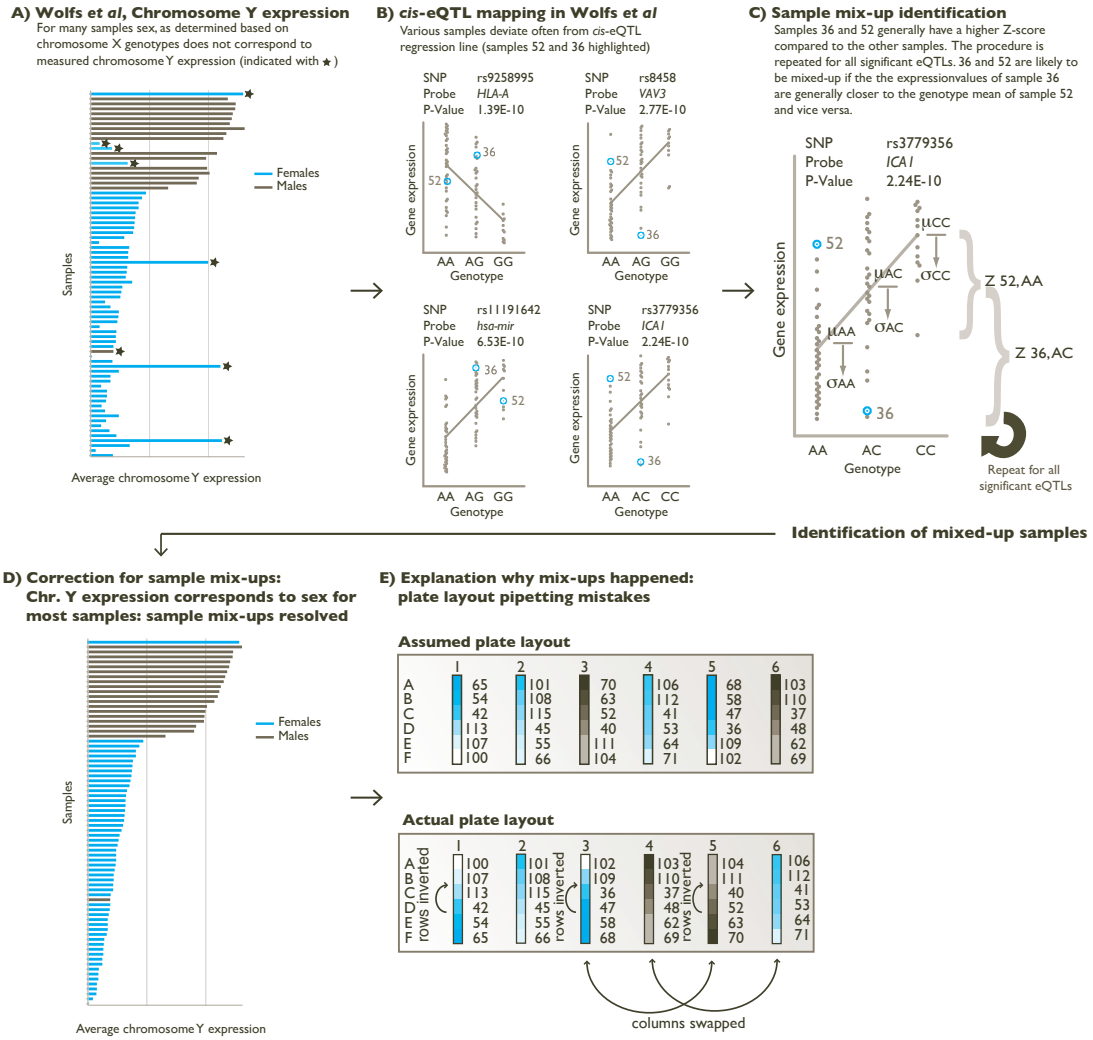


Figure 1.

A) We observed numerous sample mix-ups in a dataset created by our laboratory (Wolfs,M.G. *et al.*, unpublished data), where the chromosome Y expression did not correspond to the genotype-derived sex (indicated with asterisk). B) Four plots of *cis*-eQTLs mapped in the dataset from Wolfs,M.G. *et al.* showed samples 36 and 52 as outliers. These samples generally deviated more from the expected regression line than the other samples in this dataset (samples 36 and 52 highlighted). If this was a general observation over all significant *cis*-eQTLs for this dataset, we gathered evidence that something was wrong with these samples. C) Therefore, for each *cis*-eQTL, we calculated the mean gene expression level (μ) and standard deviation (σ) per genotype (g). This allowed us to determine, per individual (i), to what extent the gene expression level (e_i) was deviating from the regression line using an absolute Z-score$_i = |e_i - \mu_{g_i}| / \sigma_{g_i}$. Samples 36 and 52 generally have a higher Z-score compared to other samples. By repeating these calculations for all significant *cis*-eQTLs, and by comparing all pairs of gene expression arrays and genotyping arrays, we could identify those samples that were likely to be mixed-up. D) When we corrected these mix-ups we observed that the chromosome Y expression now corresponded to the sex for most samples: sample mix-ups resolved. (E) Inspection of the RNA plate layout indicated that mix-ups had been introduced by pipetting mistakes.

data. This approach was applied to all *cis*-eQTL mapping procedures in this study, unless specified otherwise.

Identifying sample mix-ups

The concept behind our method to identify sample mix-ups is straightforward: for some genes, the expression level is very strongly determined by a SNP genotype. For a pair of genotype and gene expression arrays, we can determine the concordance between the expected gene expression level conditional on the putative SNP genotype for such genes (Figure 1C). By systematically assessing all sample pairs with such a set of *cis*-eQTLs, we can determine which pairs are likely to be correct and which pairs are not. In general, there are several scenarios for errors in sample assignment. These include duplicate genotypes, duplicate expression arrays, absent genotypes, absent expression arrays and sample swaps. We considered each of these scenarios indicative of a sample mix-up. We defined the total number of sample mix-ups as the number of genotyped samples that have had an incorrect expression array assigned.

To identify sample mix-ups, MixupMapper uses each significantly detected *cis*-eQTL in the dataset. For each of these *cis*-eQTLs we determined the mean (μAA, μAB and μBB) and standard deviation (σAA, σAB and σBB) of the gene expression values for each of the three genotypes (AA, AB and BB). For this purpose we used the genotype and gene expression pairs that were initially assumed to be correct. For each pair of genotype and gene expression array we determined the SNP genotype (g). We calculated the number of standard deviations that the gene expression value (e) differed from the expected value associated with the SNP genotype using an absolute Z-score (1).

$$Z = |e - \mu_g| / \sigma_g \quad (1)$$

For each sample pair we summed the absolute Z-scores of all significant *cis*-eQTLs and determined the average Z-score for each sample pair to account for differences in the number of tested eQTLs per sample pair due to missing SNP genotypes.

Expression arrays that have been hybridized to lower quality or degraded RNA tend to result in higher deviations from the individual *cis*-eQTL regression lines, irrespective of what genotyped sample has been tested for such an expression array. As a result, such an expression sample will show higher overall Z-scores on average for each of the genotyped samples to which it is compared. Therefore, in order to standardize the Z-scores for each of the expression arrays, we normalized the Z-scores by subtracting the average of the overall Z-scores for this expression sample and divided it by the standard deviation of the overall Z-scores for this expression sample. Similarly, we normalized the Z-scores by subtracting the average of the overall Z-scores for this genotype sample and divided it by the standard deviation of the overall Z-scores for the genotype sample.

After these normalizations we determined what the expression array was with the lowest overall normalized Z-score for each genotyped sample. We considered this expression sample to reflect this particular genotyped sample.

Once the best matching expression sample had been identified for each genotyped sample, we compared it to what had been initially defined, permitting us to identify which samples were mixed-up.

It is, however, also conceivable that our method might incorrectly suggest the presence of sample mix-ups, because of potential overfitting of mean and standard deviations for each of the three genotype groups per *cis*-eQTL. We therefore used a post-hoc permutation strategy to check if our results were not due to overfitting. We first permuted the phenotype labels and subsequently ran a *cis*-eQTL analysis. As expected, this analysis did not lead to the identification of any significant *cis*-eQTL, although it did permit us to identify the list of top *cis*-eQTLs for this permutation. We chose an equal number of *cis*-eQTLs as identified in the initial *cis*-eQTL analysis (that was based on the non-permuted data). Using this set of *cis*-eQTLs and the permuted phenotype labels, we then performed the sample mix-up identification procedure. This resulted in overall Z-scores for each genotype-expression sample pair and a respective distribution that indicated what could be expected when running the mix-up procedure on randomly permuted data. Based on this distribution, we determined the 5th percentile Z-score threshold (low Z-values indicating a better agreement). We repeated this permutation strategy 1,000 times, resulting in a distribution of 5th percentile Z-score thresholds. We decided to select the 5th percentile Z-score threshold as what was attained in only 5% of the 1,000 permutations. Using this strategy we determined a Z-score significance threshold for each of the datasets. We used this threshold for each of the inferred sample mix-ups and only considered the mix-ups significant and real if the mix-up Z-score was lower than the Z-score significance threshold.

Once we had determined the best match for each genotype array, we used the following procedure to decide which genotyped samples to keep, which to correct, and which to remove completely: for each genotyped sample we first checked if the overall Z-score of the best matching gene expression array was below the permutation Z-score threshold (a lower Z-score corresponds to a better match). We discarded the genotyped sample completely if no well-matched gene expression sample was identified.

For each of the remaining genotyped samples, we checked if the best matching gene expression array corresponded to the one that was originally considered to match it. If they corresponded, this indicated that the sample had not been mixed-up and it would be kept.

For those genotyped samples with an incorrect gene expression array, we applied the following procedure: we first determined what other genotyped sample was originally coupled to that particular gene expression array. If that other genotype sample was not considered a mix-up and thus matched the particular gene expression array well, we knew we had to discard the genotype sample that we were assessing.

Alternatively, if the other genotype sample did not correlate well with the gene expression array, we knew we could safely assign the assessed genotyped sample to this gene expression array, because the other genotype sample was likely to be assigned to another well matching expression sample.

Finally, to ensure that each gene expression array was eventually assigned to a single genotype array, we checked whether there were two genotype samples that were assigned to the same gene expression sample. If this was the case, we discarded the genotype sample that was the worse match to the expression array (i.e. the one with the highest Z-score).

Using this method, we were able to not only correct for sample mix-ups, but also to identify those genotype arrays that clearly did not match any gene expression arrays. Such genotype samples generally had a worse Z-score than those genotype samples that matched a gene expression array well and they were discarded from further analyses.

Effects of sample mix-ups in genetical genomics studies

We assessed the effect of sample mix-ups in each of the assessed datasets by repeating the *cis*-eQTL mapping after we had corrected the sample mix-ups. We also simulated the effect of accidental sample mix-ups in the datasets in which we had not identified any mix-ups. The samples were permuted with increments of approximately 5%, after which we performed *cis*-eQTL mapping (FDR < 0.05, 100 permutations). We repeated this analysis 100 times to get an accurate estimate of the average number of significantly detected *cis*-eQTLs for the different sets of increasingly mixed-up samples.

Sensitivity and specificity of sample mix-up method

For the datasets where we had deliberately introduced sample mix-ups, we ascertained how many of these mix-ups could be identified by our method to get realistic estimates on its specificity and sensitivity. We defined the true positives (TP) as the number of mixed-up samples that our method had correctly identified. However, we also required that the method had identified the correct alternative expression sample. We defined the false positives (FP) as the number of samples that were either falsely deemed to be a mix-up, or that were mixed-up but not assigned to the correct alternative expression sample. We defined the true negatives (TN) as the number of samples correctly identified as not being mixed-up. This permitted us to determine the true positive rate (TPR, sensitivity) as:
TPR = TP / (Number of introduced mix-ups), and the false positive rate (FPR) as: FPR = FP / (TN + FP). We defined the specificity as 1 – FPR.

Effects of sample mix-ups on GWAS that looked at one particular, continuous, phenotype

We simulated the effect of sample mix-ups on GWAS that investigated one particular, continuous, phenotype by first generating genotypes for a population of 500,000 individuals, each with a minor allele frequency (MAF) of 0.5. On the basis of these genotypes, we then generated a random continuous

phenotype, based on an error term and the joint effect of 100, 200 or 500 unlinked SNPs (each having an equal effect size), and we assumed that each of these causal variants had been successfully genotyped. Using this method, we created phenotypes that were respectively 90%, 80%, 70%, 60% and 50% heritable. From this population, we then randomly sampled 10,000 samples, and conducted an association analysis. We correlated the phenotype to the genotype using Pearson correlation coefficients. By using a p-value threshold of 5 x 10-8, which is commonly used to declare genome-wide significance, we were able to determine what proportion of the causative variants were significant. By repeating this procedure 1,000 times, we obtained reliable estimates of how many of the variants were declared significant. We then repeated this procedure while swapping the phenotypes of an increasing number of individuals, to ascertain the effect of sample mix-ups.

Table 2.

Cis-eQTL mapping and sample mix-up identification results

Study	Population	Sample size	Initial *cis*-eQTLs	Mix-ups detected [a] *n* (%)	Sample size after correction *n* (%)	*cis*-eQTLs after correction *n* (%)
Choy et al	CHB+JPT	87	138	20 (23%)	79 (90%)	418 (+203%)
	CEU	84	558	0	N/A	N/A
	YRI	85	274	2 (2%)	83 (97%)	287 (+5%)
Stranger et al	CHB+JPT	90	1,511	0	N/A	N/A
	CEU	90	903	0	N/A	N/A
	YRI	90	663	1 (1%)	89 (99%)	667 (+1%)
Zhang et al	CEU	87	2,581	0	N/A	N/A
	YRI	89	1,454	2 (2%)	89 (100%)	1,635 (+12%)
Webster et al	Brain	363	1,284	16 (4%)	356 (98%)	1,367 (+6%)
Heinzen et al	Brain	93	349	0	N/A	N/A
	PBMC	80	297	0	N/A	N/A

[a]In four out of the five studies, sample mix-ups were present in some of the populations investigated by the authors. In a substantial number of cases, these sample mix-ups could be resolved if, for instance, we assumed that two expression samples had been accidentally swapped. Genotyped samples for which no appropriate expression sample could be identified were removed. Numbers of *cis*-eQTLs are number of unique probes with an significant effect (FDR < 0.05). N/A, not applicable.

Results

Identifying sample mix-ups

The issue of sample mix-ups became apparent in a genetical genomics dataset of liver tissue that had been generated in our lab, as there were various samples for which the chromosome Y expression did not correspond to the gender as derived from the genotypes (Wolfs *et al.*, unpublished, Figure 1A).

To identify the exact origin of these sample mix-ups we developed a sample mix-up algorithm (MixupMapper) that relies upon gene expression phenotypes that are influenced by genetic variation (*cis*-eQTLs). Fig 1B shows four *cis*-eQTL plots from this dataset in which samples 36 and 52 have been highlighted. These samples generally deviate substantially from the *cis*-eQTL regression line (Figure 1C), suggesting they have been swapped.

By applying our mix-up identification method to all pairs of genotyped and gene expression samples, we were able to identify 28 sample mix-ups in this dataset (Figure 1D). These mix-ups were later confirmed by the facility involved in generating the data: when we compared the results to the plate layout of the RNA samples used during pipetting, we observed that some columns had been swapped and some rows had been inverted after hybridization to the gene expression chip (Fig 1E). We should note that the actual mix-up of samples can have occurred during any of the steps involved in the generation of this dataset, such as during DNA and RNA isolation, aliquoting, and hybridization. However, these samples appeared to be mixed-up because of pipetting mistakes prior to hybridization on the RNA-chip, as we did not observe any indications of errors that occurred in the DNA preparation or hybridization process. In total, 30 out of 74 samples (41%) had been assigned wrong expression phenotypes and our method was able to resolve 28 of them.

Sample mix-ups in published datasets

As we had identified these mix-ups in our own data, we applied our method to five publicly available human datasets for which both genotype and gene expression data was freely available online (Table 1)[18–22]. Four out of these five datasets contained sample mix-ups (Table 2, Figure S1, Table S1) and observed that for 3% of all samples, the genotype and expression data did not correspond (41 out of 1,238 samples). The number of sample mix-ups was highest in the CHB+JPT subset from Choy et al: out of 87 samples, 20 were incorrect (23%). In total, we were able to correct 21 of the 41 samples that had an incorrect expression phenotype.

We assessed whether the identified mix-ups were not due to extreme expression levels (e.g. because of hybridization problems that lead to poor normalization) or bad quality genotypes. For this purpose, we assessed the variability of all genotype and gene expression samples within each dataset using principal component analysis (PCA) on the sample correlation matrix: samples that are clear outliers in terms of sample quality, show deviations for the first two principal components (PCs). To test whether the identified sample mix-ups deviated from the remaining samples, we performed a Wilcoxon-Mann-Whitney test on the eigenvalues for these PCs. As we did not observe any significant differences

(Bonferroni correction P<0.004), we can conclude that sample quality does not confound the results of our method (Fig S2).

Sensitivity and specificity

We established that our method is highly specific and sensitive, by conducting simulations in the datasets in which no mix-ups had been identified (Figure 2A). We observed the best performance in the Stranger *et al.* CHB+JPT dataset: even when 40% of the samples had been mixed-up, 98% (σ^2: 3%) of these 'errors' could still be successfully identified and successfully corrected. The worst performance was observed in the Heinzen *et al.* post-mortem brain dataset: when 10% of the samples in this dataset were mixed-up, only 85% (σ^2: 15%) of these could be successfully identified and corrected. Very few samples were wrongly deemed a sample mix-up by our method, indicating our method is highly specific: for each of the datasets in which 10% of the samples had been mixed-up, we observed only rarely that non-mixed-up samples were wrongly deemed to be mixed-up (Specificity > 99.9%, Figure 2A).

Effect of correcting sample mix-ups on number of detected *cis*-eQTLs

Table 2 shows that the number of detected *cis*-eQTLs increased in each of the datasets after we had corrected for the sample mix-ups (Table 2, and Figure S3, Figure S4A-S4F, FDR < 0.05, 1,000 permutations). In total 15% more *cis*-eQTLs were identified for these datasets. For the CHB+JPT population from Choy et al,

24
Benovoy, D., Kwan, T. & Majewski, J. Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. *Nucleic Acids Res.* 36, 4417–23 (2008).

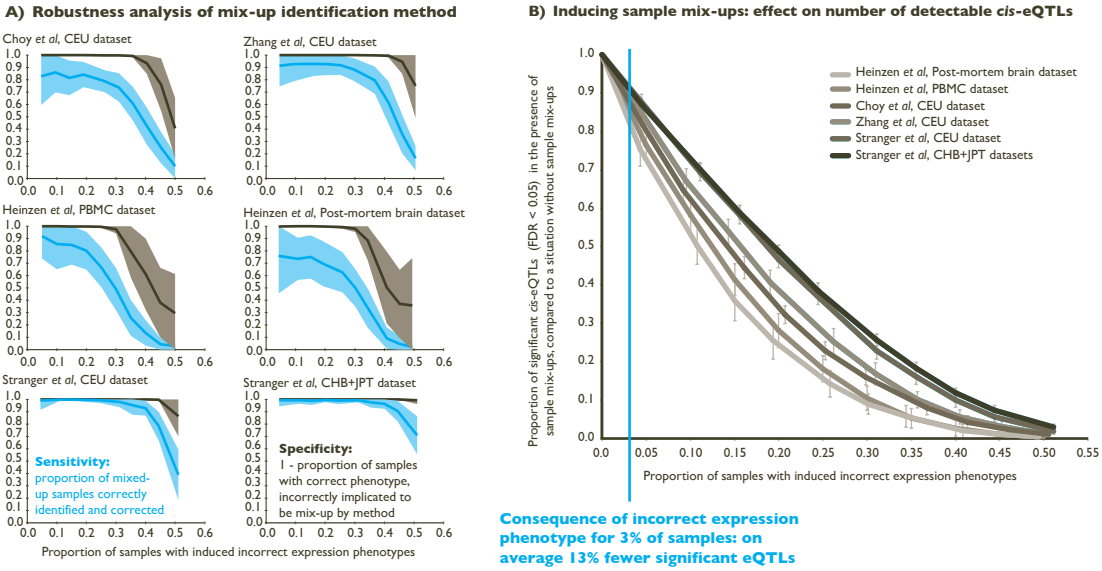


Figure 2.
A) Robustness analyses on the sample mix-up identification method shows that if the fraction of mixed-up samples is <25%, most of the sample mix-ups are detected and corrected by our method with high specificity and sensitivity. Variability is generally low (shaded areas around graphs), especially for the specificity. B) Sample mix-ups were introduced into the six datasets that did not initially contain any mix-ups. Deliberately introducing sample mix-ups resulted in a substantial decrease in the number of significantly detectable *cis*-eQTLs (FDR < 0.05). If 3% of the samples had an incorrect phenotype assigned, the average number of detectable *cis*-eQTLs decreased by 13%.

correcting the identified mix-ups increased the number of significant *cis*-eQTLs by 203% (418 *cis*-eQTLs compared to 138 *cis*-eQTLs before correction). This is a considerable increase, especially since the effective sample size decreased by 9% (79 samples included after correction compared to 87 samples in the original dataset). Furthermore, this indicated that the removed samples effectively amounted to noise and therefore contributed to a decrease in the power to detect *cis*-eQTL effects, especially smaller ones. However, the increase in the number of detected *cis*-eQTLs we describe here could also be explained by an increase of the proportion of false positives that are the result of SNPs being present in the gene expression probe sequences, directly affecting hybridization efficiency[24]. We therefore checked whether the proportion of potential false positives due to probe polymorphisms differed before and after sample mix-up correction for each dataset and found no differences in the assessed datasets (Figure S5).

Replication of detected *cis*-eQTLs

We reasoned that we could gain further evidence that the identified sample mix-ups were indeed mix-ups by replicating the *cis*-eQTLs that were identified after the correction. It has been shown that substantial overlap exists in *cis*-eQTL effects between different populations and also between different tissues[20,25,26]. We therefore compared the datasets that had been run on the same expression platform, which amounts to comparing the different HapMap populations. For the studies of Choy *et al.*, Cox *et al.* and Stranger *et al.*, we assessed how many of the significantly detected *cis*-eQTLs in one HapMap population had also been detected in another HapMap population (Figure S6). After correcting the sample mix-ups, the number of shared *cis*-eQTLs increased in each of the population comparisons. 99.7% of the eQTLs that were shared between at least two populations showed identical allelic directions. We observed comparable increases in shared *cis*-eQTLs after mix-up correction, when comparing identical HapMap populations that had been run on different expression platforms (Figure S7).

Effect of sample mix-ups on GWAS studies on continuous traits

We systematically explored the effect of different proportions of sample mix-ups on the power to detect *cis*-eQTLs. We investigated the datasets in which we had not identified any sample mix-ups and deliberately introduced sample mix-ups by swapping the expression array measurements for an increasing number of samples. We observed a considerable decrease of, on average, 13% when only 3% of the samples were deliberately mixed-up (Figure 2B).

Furthermore, we decided to model the influence of sample mix-ups on GWAS studies that investigate traits caused by hundreds of variants, such as height or body mass index. We simulated a trait with several degrees of heritability and different numbers of causative variants in a population of 500,000 individuals. From this population we randomly sampled 10,000 individuals on which we conducted a linear regression analysis and determined the percentage of SNPs that were genome-wide significant ($P < 5 \times 10\text{-}8$, Figure 3A). To measure the effect of

25
Bullaughey, K., Chavarria, C. I., Coop, G. & Gilad, Y. Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Hum. Mol. Genet.* 18, 4296–303 (2009).
26
Heap, G. A. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* 2, 1 (2009).

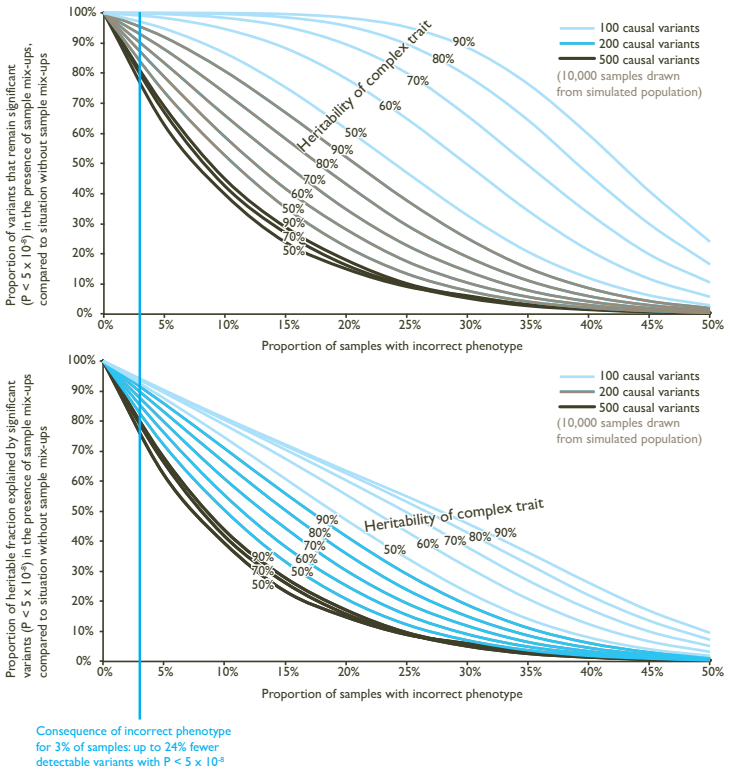
sample mix-ups, we then deliberately randomized the trait measurements of an increasing number of randomly selected individuals (Figure 3B and Figure 3C).

A very limited number of sample mix-ups had severe consequences for traits that have a heritability of 50%, of which the heritable fraction is due to 500 causal variants. If 3% of all samples had an incorrect phenotype (as observed in the genetical genomics datasets), we could only significantly detect 77% of the causal variants that we would have identified if no mix-ups had been present (at P < 5 x 10-8, Figure 3B). Likewise, we observed a decrease in the proportion of the heritability that could be explained by the genome-wide significant associations (Figure 3C): the explained heritability of genome-wide significant variants decreased with 1.24 fold (from 3.7% to 2.8%) when 3% of the samples were mixed-up.

A) Results from linear regression analysis in 10,000 samples, drawn from simulated population if no sample mix-ups exist

Number of causal variants	Heritability of trait	Proportion of causal variants with association P < 5 x 10^{-8}	Prop. heritability explained by causal variants with P < 5 x 10^{-8}
100	0.9	100%	100%
	0.8	100%	100%
	0.7	100%	100%
	0.6	99%	99%
	0.5	95%	97%
200	0.9	90%	94%
	0.8	81%	89%
	0.7	68%	80%
	0.6	51%	65%
	0.5	33%	48%
500	0.9	11%	22%
	0.7	4%	10%
	0.5	1%	4%

B) Loss in proportion of genome-wide signifcant causal variants in the presence of sample mix-ups



C)

Figure 3.

A) Robustness analyses on the sample mix-up identification method shows that if the fraction of mixed-up samples is <25%, most of the sample mix-ups are detected and corrected by our method with high specificity and sensitivity. Variability is generally low (shaded areas around graphs), especially for the specificity. B) Sample mix-ups were introduced into the six datasets that did not initially contain any mix-ups. Deliberately introducing sample mix-ups resulted in a substantial decrease in the number of significantly detectable *cis*-eQTLs (FDR < 0.05). If 3% of the samples had an incorrect phenotype assigned, the average number of detectable *cis*-eQTLs decreased by 13%.

If these sample mix-ups could be detected and corrected, it would be possible to explain 1.24 fold more of the heritability for such traits. Methods to detect such sample mix-ups therefore have the potential to substantially increase the explained heritability and power to detect genetic effects in GWAS on complex phenotypes such as height, BMI or some diseases.

Discussion

We have identified sample mix-ups in four out of five genetical genomics studies by applying a novel method (MixupMapper). On average, 3% of all samples were mixed-up. After correction for these sample mix-ups by our method, we detected on average 15% more *cis*-eQTLs. Correcting mix-ups in one dataset in which 23% of the samples were incorrect led to three times as many significant *cis*-eQTLs being detected. The consequences of only 3% sample mix-ups on the heritable fraction that can be explained by significantly associated variants can also be substantial. For some simulated complex traits with a moderate to high heritability, `the explained heritability of the genome-wide significant variants increased 1.24 fold, when these sample mix-ups could be detected and corrected.

A considerable proportion of the heritability of complex diseases and traits is currently 'missing'. There is debate on whether the missing heritability problem is caused by rare variants with a large effect, by many more common variants, each with a very small effect size, by overestimation of the heritability estimates or through other means[17,27]. As current genome-wide studies are pushing towards associating ever smaller effect sizes, sample sizes have to increase substantially to discover loci with smaller effect sizes[28]. Our results indicate that especially for these small effect size loci, sample mix-ups could have consequences on the power to detect such loci for both genetical genomics studies as well as GWAS. As such, a proportion of the missing heritability could possibly be explained by the presence of sample mix-ups in genome-wide datasets.

However, it remains a question whether the frequency of sample mix-ups we observed in genetical genomics samples is a realistic estimate for other types of genome-wide datasets. Different types of genome-wide datasets require many different handling steps, and therefore their frequencies for the presence of sample mix-ups may differ. For example for case-control GWAS the cases and controls are often collected and processed separately from each other. Therefore, the frequency and hence the consequence of sample mix-ups in such case-control studies might be lower compared to the studies presented here. As such, GWAS in general might contain fewer sample mix-ups.

In the case of genetical genomics datasets, more *cis*-eQTLs could be detected in each of the datasets after correction, although the number of included samples had actually decreased for three of these datasets. This effectively demonstrates that increasing the sample size is not the only way of increasing statistical power for determining complex traits; increasing the phenotypic accuracy can be equally helpful. In addition to the

27
McCarthy, M. I. & Hirschhorn, J. N. Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* 17, R156–65 (2008).
28
Park, J.-H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* 42, 570–5 (2010).
29
Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–35 (2007).

method described here, phenotypic accuracy may be increased by, for example, including relevant co-variates in GWAS, or by using principal components analysis in *cis*-eQTL studies[2,29]. Although these methods are helpful in increasing the phenotypic accuracy, they do not help to identify or overcome sample mix-ups.

One possible problem with MixupMapper is that it depends on an initial set of *cis*-eQTLs that can be detected in the data. Although our method generally shows high sensitivity and specificity when large proportions of samples are mixed-up, in an extreme scenario, where all genotyped samples are randomly assigned to the gene expression samples, the mix-ups cannot be resolved since no *cis*-eQTLs will be initially detectable. However, if a set of *cis*-eQTLs has been independently identified in another set of samples for the particular expression platform used, it is also be possible to resolve these problems (data not shown).

We feel it is important to emphasize that the sample mix-ups that we detected in the five public datasets do not in any way discredit these studies. To our knowledge, we are the first to describe a method to identify mix-ups for these kinds of datasets. If the authors had been aware of the existence of these mix-ups, they would have certainly corrected them, as their goal was to find as many eQTLs as possible. Since we observed a substantial overlap of detected *cis*-eQTLs in the three HapMap populations before correction of sample mix-ups, we assume the detected *cis*-eQTLs in these datasets are still valid. We are convinced that the results and conclusions drawn from these datasets[18–22] remain appropriate.

Although our method is intended for genetical genomics datasets, it can also be applied to other types of genome-wide datasets, as long as there are sufficient numbers of different phenotypes available per individual that are each (strongly) determined by genetic variants or combinations of variants. This requirement will likely be met with the growing interest in population-based cohort studies in which hundreds of phenotypes are collected from the participants. As a consequence, identifying sample mix-ups will then become possible for these datasets as well.

Our results clearly indicate that sample mix-ups occur in many labs (including ours). Although a great deal of quality control is conducted in GWAS, it is very difficult to prevent the accidental mislabeling of some samples. This is particularly problematic in studies of unrelated individuals where inheritance patterns cannot be investigated. Nevertheless these accidental human mistakes or experimental problems can sometimes have far-reaching consequences. We therefore recommend rigorous quality control of the laboratory and administrative processes in order to prevent sample mix-ups from happening. We conclude that fewer sample mix-ups will increase the power to detect significant genetic associations substantially and might resolve a part of the missing heritability.

Acknowledgements

We thank Jackie Senior for her critical reading of the manuscript. Funding: This work was supported by a Horizon Breakthrough grant from the Netherlands Genomics Initiative (93519031), a VENI grant from the Netherlands Organization for Scientific Research (NWO, ZonMW grant 916.10.135) and has received funding from the European Community's Health Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 259867 to L.F.

Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA

PLoS Genetics, 2011 August; 7, 14

Rudolf S. N. Fehrmann[1], Ritsert C. Jansen[2,], Jan H. Veldink[3,], Harm-Jan Westra[1,], Danny Arends[2], Marc Jan Bonder[1], Jingyuan Fu1, Patrick Deelen[1], Harry J. M. Groen[4], Asia Smolonska[1], Rinse K. Weersma[1,5], Robert M. W. Hofstra[1], Wim A. Buurman[6], Sander Rensen[6], Marcel G. M. Wolfs[7], Mathieu Platteel[1], Alexandra Zhernakova[8], Clara C. Elbers[9], Eleanora M. Festen[1], Gosia Trynka[1], Marten H. Hofker[7], Christiaan G. J. Saris[3], Roel A. Ophoff[3,10,11], Leonard H. van den Berg[3], David A. van Heel1[2], Cisca Wijmenga[1], Gerard J. te Meerman[1"], Lude Franke[1,12*"]

3

1 Department of Genetics, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands
2 Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Haren, The Netherlands
3 Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Utrecht, The Netherlands
4 Department of Pulmonology, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands
5 Department of Gastroenterology and Hepatology, University Medical Centre Groningen and University of Groningen, Groningen, The Netherlands
6 NUTRIM School for Nutrition, Toxicology, and Metabolism, Department of General Surgery, Maastricht University Medical Center, Maastricht, The Netherlands
7 Department of Pathology and Medical Biology, Medical Biology Section, Molecular Genetics, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands
8 Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands
9 Departments of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America
10 Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands
11 Center for Neurobehavioral Genetics, University of California Los Angeles, Los Angeles, California, United States of America
12 Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom

Author Summary

Many genetic variants have been found associated with diseases. However, for many of these genetic variants, it remains unclear how they exert their effect on the eventual phenotype.
We investigated genetic variants that are known to be associated with diseases and complex phenotypes and assessed whether these variants were also associated with gene expression levels in a set of 1,469 unrelated whole blood samples. For several diseases, such as type 1 diabetes and ulcerative colitis, we observed that genetic variants affect the expression of genes, not implicated before. For complex traits, such as mean platelet volume and mean corpuscular volume, we observed that independent genetic variants on different chromosomes influence the expression of exactly the same genes. For mean platelet volume, these genes include well-known blood coagulation genes but also genes with still unknown functions. These results indicate that, by systematically correlating genetic variation with gene expression levels, it is possible to identify downstream genes, which provide important avenues for further research.

Abstract

For many complex traits, genetic variants have been found associated. However, it is still mostly unclear through which downstream mechanism these variants cause these phenotypes. Knowledge of these intermediate steps is crucial to understand pathogenesis, while also providing leads for potential pharmacological intervention. Here we relied upon natural human genetic variation to identify effects of these variants on *trans*-gene expression (expression quantitative trait locus mapping, eQTL) in whole peripheral blood from 1,469 unrelated individuals.
We looked at 1,167 published trait- or disease-associated SNPs and observed *trans*-eQTL effects on 113 different genes, of which we replicated 46 in monocytes of 1,490 different individuals and 18 in a smaller dataset that comprised subcutaneous adipose, visceral adipose, liver tissue, and muscle tissue. HLA single-nucleotide polymorphisms (SNPs) were 10-fold enriched for *trans*-eQTLs: 48% of the *trans*-acting SNPs map within the HLA, including ulcerative colitis susceptibility variants that affect plausible candidate genes *AOAH* and *TRBV18* in *trans*. We identified 18 pairs of unlinked SNPs associated with the same phenotype and affecting expression of the same *trans*-gene (21 times more than expected, $P<10^{-16}$). This was particularly pronounced for mean platelet volume (MPV): Two independent SNPs significantly affect the well-known blood coagulation genes *GP9* and *F13A1* but also *C19orf33*, *SAMD14*, *VCL*, and *GNG11*. Several of these SNPs have a substantially higher effect on the downstream *trans*-genes than on the eventual phenotypes, supporting the concept that the effects of these SNPs on expression seems to be much less multifactorial. Therefore, these *trans*-eQTLs could well represent some of the intermediate genes that connect genetic variants with their eventual complex phenotypic outcomes.

Table 1.

Detected eQTLs in 1,469 genetical genomics samples for 289,044 common SNPs and for 1,167 trait-associated SNPs.

eQTL analysis on 289,044 common SNPs	***cis*-eQTLs (FDR<0.05)**	***trans*-eQTLs (FDR<0.05)**
Spearman's correlation threshold	$P < 1.73 \times 10^{-3}$	$P < 3.6 \times 10^{-9}$
Number of tests performed	2,329,207	13,292,122,142
Number of unique eQTL probes	10,872	244
Number of unique eQTL genes	7,589	202
Number of unique eQTL SNPs	48,717 (16.9% of all tested SNPs)	467 (0.2% of all tested SNPs)
Number of unique MHC eQTL SNPs	1,586 (3.3% of *cis*-eQTL SNPs)	155 (33.2% of *trans*-eQTL SNPs)

eQTL analysis on 1,167 trait-associated SNPs	***cis*-eQTLs (FDR<0.05)**	***trans*-eQTLs (FDR<0.05)**	***trans*-eQTLs (FDR<0.5)**
Spearman's correlation threshold	$P < 3.7 \times 10^{-3}$	$P < 2.0 \times 10^{-7}$	$P < 1.02 \times 10^{-5}$
Number of tests performed	15,371	53,629,458	53,629,458
Number of unique eQTL probes	679	130	726
Number of unique eQTL genes	538	113	576
Number of unique eQTL SNPs	472 (40.4% of all tested SNPs)	67 (5.7% of all tested SNPs)	462 (39.6% of all tested SNPs)
Number of unique MHC eQTL SNPs	65 (13.8% of *cis*-eQTL SNPs)	32 (47.8% of *trans*-eQTL SNPs)	52 (11.3% of *trans*-eQTL SNPs)

For 289,044 SNPs, present on the commonly used Illumina HumanHap300 platform, the false discovery rate (FDR) was controlled at 0.05 for both *cis*- and *trans*-eQTLs. For the analysis of 1,167 successfully imputed SNPs that have been found associated with a quantitative trait or disease the FDR was controlled at 0.05 for the *cis*- and *trans*-eQTLs. We also performed a *trans*-eQTL analysis for these SNPs while controlling the FDR at 0.50 to generate more hypotheses. The number of unique genes was determined using Ensembl 52 (NCBI 36.3 release).

Introduction

For many complex traits and diseases, numerous associated single nucleotide polymorphisms (SNPs) have been identified through genome-wide association studies (GWAS) through genome-wide association studies (GWAS)[1]. For many of these identified variants it is still unclear through which mechanism the association between the SNP and the trait or disease phenotype is mediated. A complicating factor is that disease-associated variants might not be the real causal variants, but are in linkage disequilibrium (LD) with the true disease-causing variant, making it difficult to accurately implicate the correct gene for a locus in disease pathogenesis.

Within the major histocompatibility locus (MHC) on 6p, many SNPs have been found to be associated with complex diseases such as celiac disease, inflammatory bowel disease, psoriasis,

1 Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9362–7 (2009).

2 Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302 (2010).

3 Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* 41, 703–7 (2009).

4 Gregersen, P. K. *et al.* REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat. Genet.* 41, 820–3 (2009).

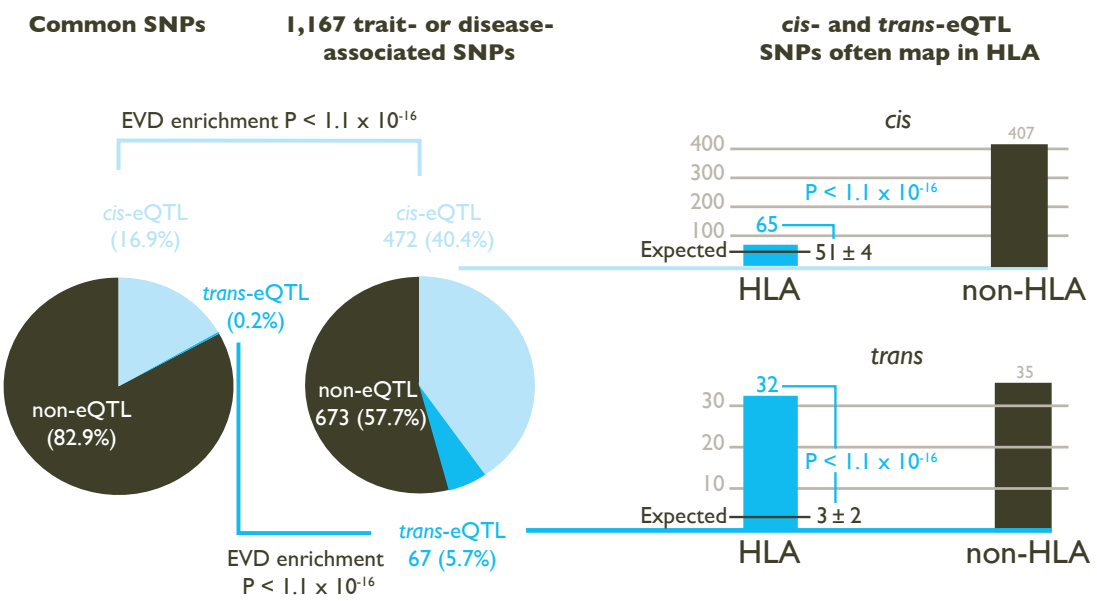


Figure 1. Disease and trait-associated SNPs are enriched for both *cis*- and *trans*-eQTLs.

17% of SNPs, present on common SNP platforms, affect gene expression levels in *cis* or *trans* (at FDR of 0.05). This is substantially different from 1,167 SNPs that have been found associated with traits or disease: 40.4% affect gene expression in *cis*, while 5.7% of these SNPs affect gene expression in *trans*. These eQTL SNPs significantly more often than expected map within the HLA (13.8% of *cis*-eQTLs, 47.8% of *trans*-eQTLs, extreme value distribution $p < 1 \times 10^{-16}$).

rheumatoid arthritis, diabetes mellitus, schizophrenia, lung cancer and follicular lymphoma[2–10]. An analysis of the Catalog of Published Genome-Wide Association Studies[1] revealed that out of 1,167 unique SNP associations with a reported $p < 5 \times 10^{-7}$, 82 (7.0%) were located within the MHC (Fisher's Exact $p < 10^{-30}$). Except for celiac disease[11] it remains largely unclear how MHC variants increase disease susceptibility.

However, common variants have been identified that might exert their function by altering gene expression rather than by altering protein structure[2,12–16] (expression quantitative trait loci, eQTLs). Comprehensive eQTL mapping (or genetical genomics[17]) will enable us to assess for every known disease-associated variant if it significantly affects gene expression. Genetic variants that affect expression of genes that map in their vicinity (*cis*-eQTLs) can potentially pinpoint the true disease gene from an associated locus. In addition, genetic variants may also affect expression of genes that reside further away or are on different chromosomes (*trans*-eQTLs)[18]. These *trans*-eQTLs are especially interesting, since they allow us to identify downstream affected disease genes which were not implicated by GWAS studies at all, and thereby potentially having the ability to reveal previously unknown (disease) pathways.

In this study we performed a comprehensive eQTL mapping to explore the downstream effects of SNPs on gene expression by analyzing genotype and expression data of 1,469 unrelated samples. In addition to a genome-wide analysis, we also performed a

5 Imielinski, M. *et al.* Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.* 41, 1335–40 (2009).

6 Kochi, Y. *et al.* A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. *Nat. Genet.* 42, 515–9 (2010).

7 Nair, R. P. *et al.* Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.* 41, 199–204 (2009).

8 O'Donovan, M. C. *et al.* Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat. Genet.* 40, 1053–5 (2008).

9 Skibola, C. F. *et al.* Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat. Genet.* 41, 873–5 (2009).

10 Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* 40, 1407–9 (2008).

11 Van Heel, D. A. & West, J. Recent advances in coeliac disease. *Gut* 55, 1037–46 (2006).

12 Choy, E. *et al.* Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* 4, e1000287 (2008).

13 Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* 39, 1202–7 (2007).

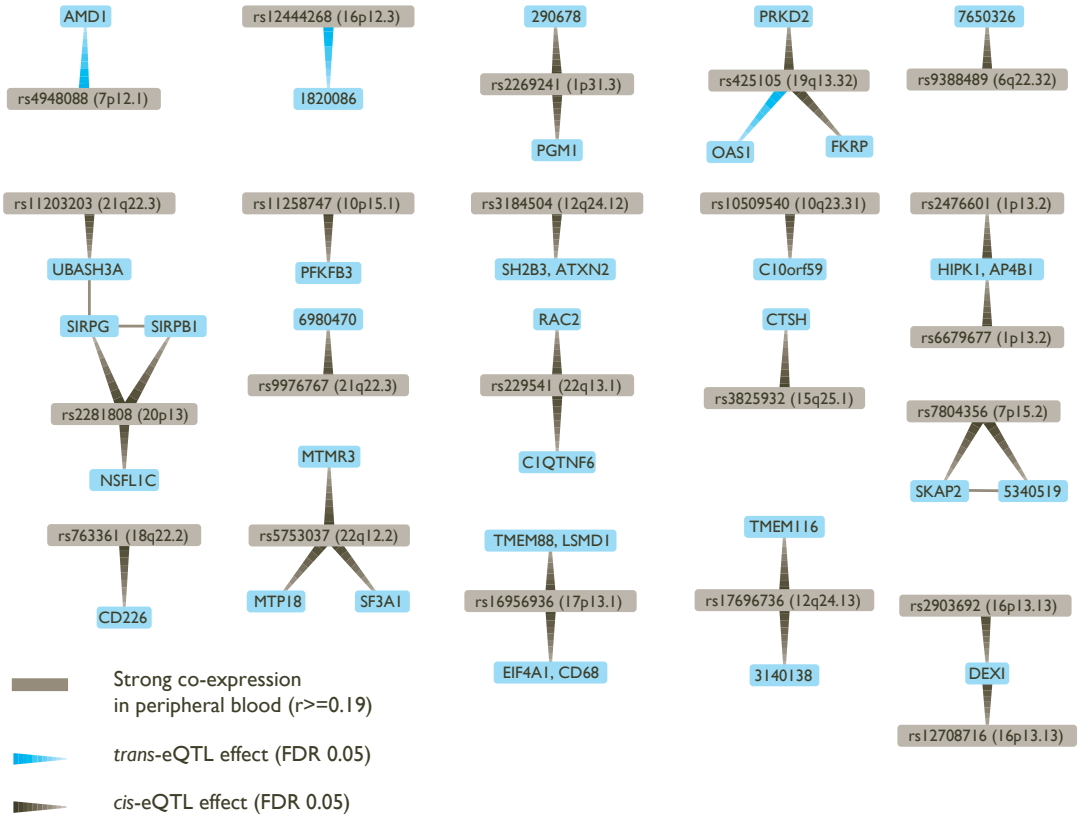


Figure 2. Type 1 diabetes associated SNPs both affect genes in *cis* and *trans*

focused analysis for disease-and trait-associated SNPs and SNPs located within the HLA. We replicated the identified *trans*-eQTLs in a collection of monocyte expression data and expression data from subcutaneous adipose, visceral adipose, muscle and liver tissue. Principal component analysis (PCA) enabled us to remove non-genetic expression variation[19,20], resulting in increased power to detect eQTLs. A stringent probe-mapping strategy was used to filter out false-positive *cis*-eQTLs due to primer-polymorphisms and false-positive *trans*-eQTLs due to cross-hybridizations. Furthermore, a permutation strategy was utilized that corrects for multiple-testing, while preventing potential confounders such as non-even distribution of SNP markers and expression probe markers across the genome, differences in minor allele frequency (MAF) between SNPs, linkage disequilibrium (LD) within the genotype data, and correlation between expression probes.

Results

Cis- and *trans*-eQTL mapping

Results of a genome-wide eQTL analysis on 289,044 common SNPs, present on the Illumina HumanHap300 platform in peripheral blood expression data of 1,469 unrelated individuals, are provided in Table 1, Table S1, Table S2, Figure S1 (controlling false discovery rate (FDR) at 0.05 using a permutation strategy).

14 Heap, G.A. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* 2, 1 (2009).

15 Moffatt, M. F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448, 470–3 (2007).

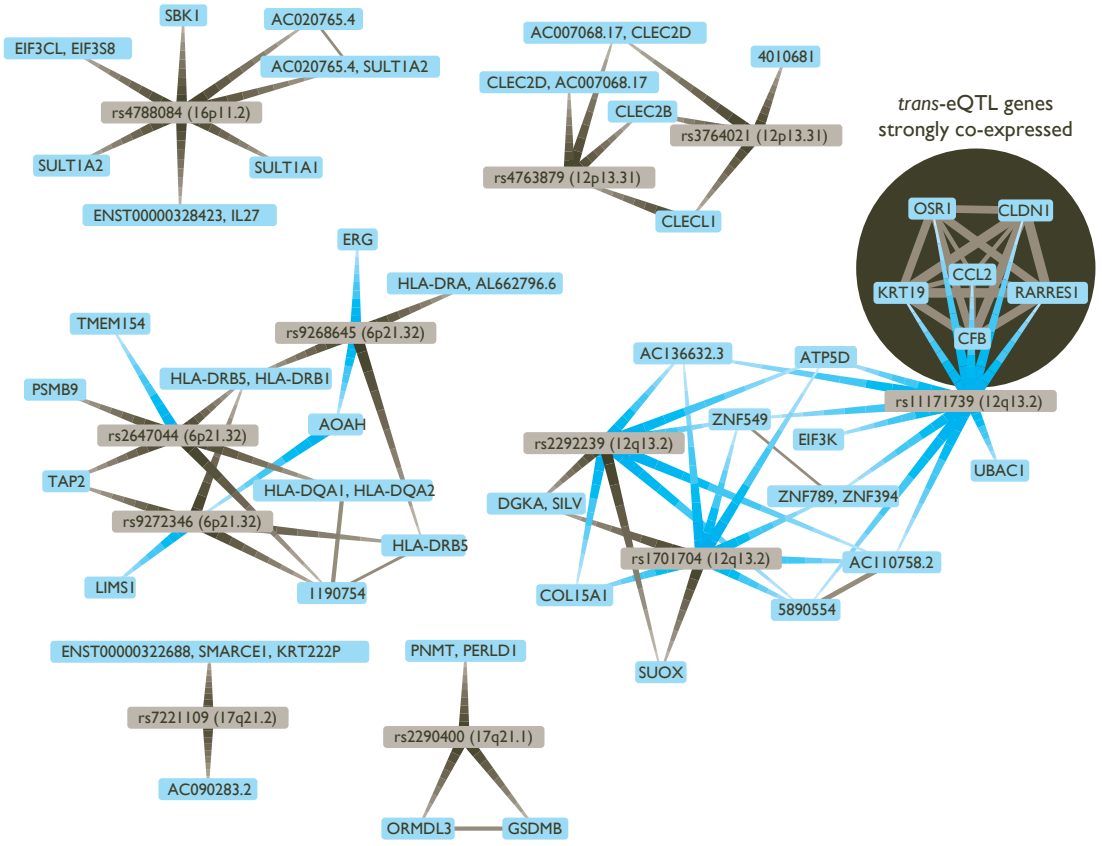
16 Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848–53 (2007).

17 Jansen, R. C. & Nap, J. P. Genetical genomics: the added value from segregation. *Trends Genet.* 17, 388–91 (2001).

18 Idaghdour, Y. *et al.* Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat. Genet.* 42, 62–7 (2010).

19 Biswas, S., Storey, J. D. & Akey, J. M. Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics* 9, 244 (2008).

20 Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–35 (2007).



As reported before[21–25] we also observed that eQTLs are strongly enriched for trait-associated SNPs (SNPs associated with a trait or disease phenotype, as reported in the Catalog of Published Genome-Wide Association Studies[1]): We therefore concentrated on these variants and imputed (Impute v2.026) additional genotype data permitting us to test 1,167 trait-associated SNPs. After removing false-positive eQTLs due to primer-polymorphisms and cross-hybridization 472 (40.4%) of these SNPs were *cis*-eQTLs, affecting the expression of 679 different transcripts, representing 538 genes (Figure 1, Table 1, Figure S2, Table S3).

67 (5.7%) SNPs were *trans*-acting on 130 different transcripts, representing 113 genes (Table S4). Results on the number of detected eQTLs per complex trait are provided in Table S5 and Figure S3. For nearly all significant *trans*-eQTLs the effect was present in each of the seven individual patient and controls cohorts, making up the total dataset (Table S6).

These *trans*-eQTLs provide valuable insight on previously unknown functional downstream consequences trait-associated SNPs have, e.g. rs2395185 is the strongest susceptibility variant for ulcerative colitis[27] (UC) but also the strongest SNP, *trans*-acting on Acyloxyacyl hydrolase (*AOAH*, $p = 1.0 \times 10^{-36}$), an enzyme that modulates host inflammatory responses to gram-negative bacterial

invasion. It is known that deficiencies in response mechanisms against bacterial products like lipopolysaccharide, present on gram-negative bacterial cell walls, play an important role in UC disease pathogenesis[28]. Within the peripheral blood we observed that *AOAH* is significantly co-expressed with colony stimulating factor 1 receptor (*CSF1R*, r = 0.21) and major histocompatibility complex class II DR alpha (*HLA-DRA*, r = 0.19). Hyperstimulation of *CSF1R* has been implicated in UC[29], while *HLA-DRA* is one of the positional UC candidate genes mapping in very close proximity to rs2395185. Another UC HLA variant, rs9268877, was *trans*-acting on T cell receptor beta variable 18 (*TRBV18*), part of the TCRß locus at 7q34. It is known that TCRß mutant mice develop chronic colitis[30].

21
Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6, e1000895 (2010).

22
Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6, e1000888 (2010).

23
Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* 452, 423–8 (2008).

24
Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6, e107 (2008).

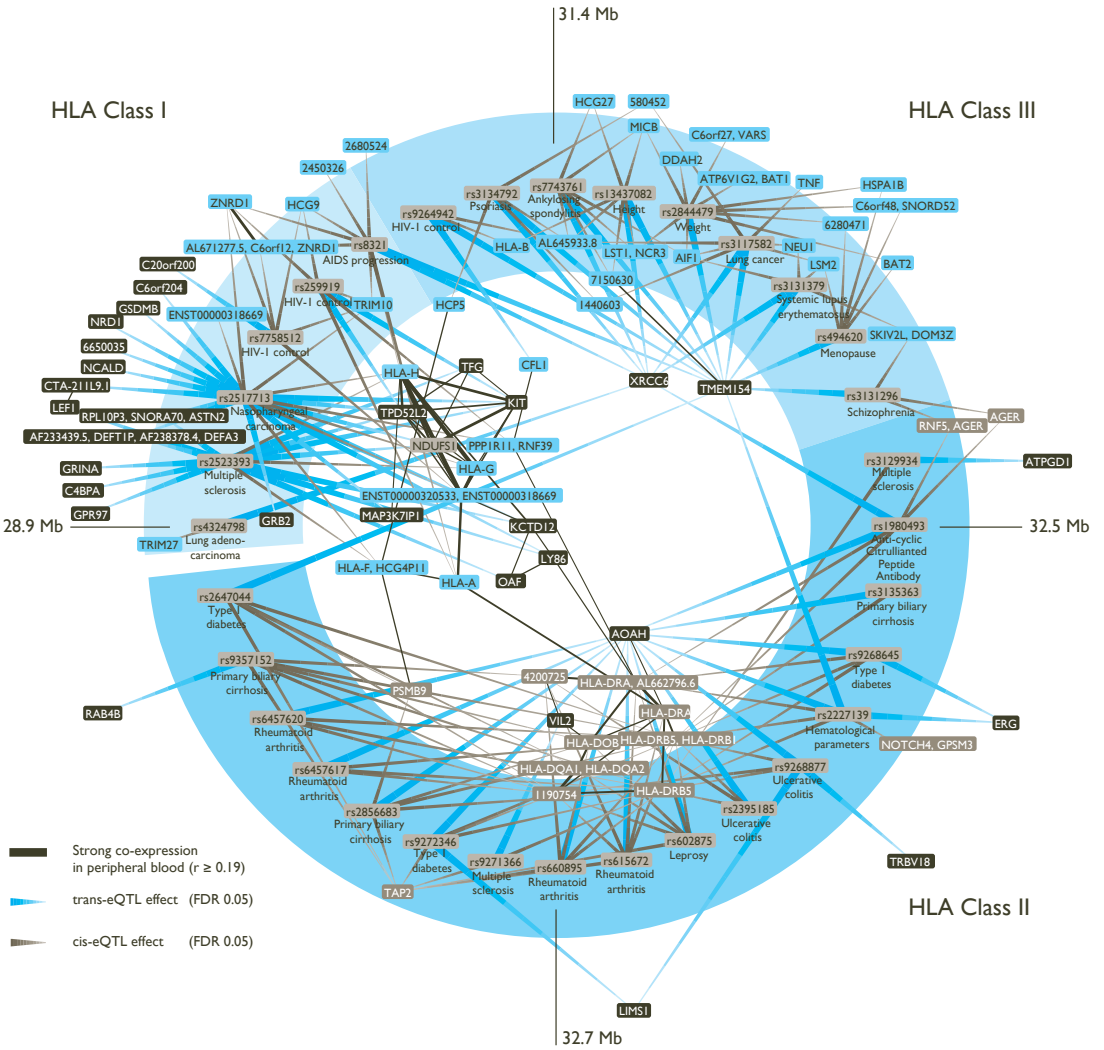


Figure 3. Human leukocyte antigen (HLA) trait-associated SNPs affect gene expression levels in *trans*.

Thirty-two trait-associated SNPs that map within the HLA are *trans*-acting on other genes. *Trans*-genes are indicated in red. Peripheral blood co-expression (Pearson correlation coefficient r≥0.19, $p<10^{-11}$) between genes is indicated in light grey. Several *trans*-genes are co-expressed with HLA genes.

For type 1 diabetes (T1D) we observed that 59% (30/51) of the known and tested T1D associated SNPs are *cis*-acting (on in total 53 unique genes) and 17% (9/50) are *trans*-acting on 22 unique genes (Figure 2). Potentially interesting *trans*-genes include *CCL2, CFB, CLN1, KRT19, OSR1* and *RARRES1*, all strongly co-expressed with each other. *CCL2* and *CFB* are known immune response genes and have been implicated in T1D before[31–33]. For breast cancer we observed that rs3803662[34] is *trans*-acting on origin recognition complex subunit 6 (*ORC6L*). This gene is involved in DNA replication and has been frequently used as part of prognostic profiles for predicting the clinical outcome in breast cancer[35,36].

We observed a marked enrichment for SNPs within the MHC among the *cis*- and *trans*-acting trait-associated SNPs: 65 of 472 *cis*-acting SNPs (13.8%, EVD $p < 1.0 \times 10^{-16}$) and 32 of 67 *trans*-acting SNPs (47.8%, EVD $p < 1.0 \times 10^{-16}$) mapped within the MHC (Figure 3). These SNPs all map to the Human Leukocyte Antigens (HLA) locus: SNPs within the HLA class I region, class II region and class III region affect 20, 7 and 2 different genes in *trans*, respectively.

Biological convergence of *cis*- and *trans*-eQTLs

While multiple associated SNPs have been identified for many complex diseases, it often remains unclear what the intermediate effects of these variants are that eventually lead to disease. It is reasonable to assume that for a particular phenotype the different associated SNPs eventually converge on the same downstream gene(s) or pathways.

We identified 7 unique pairs of unlinked SNPs that are associated with the same phenotype and that also affect the same downstream genes in *trans* or *cis* (at FDR 0.05, Table 2, Figure 4A). In order to establish whether this was more than expected by chance, we repeated this analysis, while using a set of *trans*-eQTLs, equal in size to the set of real *trans*-eQTLs, most significant after having permuted the expression sample identifiers. We performed this procedure 100 times, and observed on average only 0.15 unique pairs of unlinked SNPs (range [0, 3], Figure 4B) that showed this convergence, which indicates that the observed number of converging pairs of SNPs is 47 times more than expected (EVD $p < 1.0 \times 10^{-16}$) and implies a false-positive rate of 0.021.

Due to this highly significant enrichment of converging pairs of SNPs and its low estimated false-positive rate, we also ran an analysis where we had relaxed the FDR for *trans*-eQTLs to 0.50 (Table S7). Here we observed 18 pairs of SNPs that converge on the same genes, whereas in the 100 subsequent permutations we observed this only on average for 0.84 SNP-pairs (range [0, 5], 21 times more expected by chance, EVD $p < 1.0 \times 10^{-16}$, implying a false-positive rate of 0.047, Table 2, Figure 4B).

Many of these converging downstream genes make biological sense: three independent loci, associated with hemoglobin protein levels[37–39] and β thalassemia susceptibility[40], significantly affect hemoglobin gamma G (*HBG2*) gene expression levels (each with $p < 1.0 \times 10^{-23}$, Figure 5). For mean corpuscular volume (MCV, Figure 5) two unlinked MCV SNPs[41,42] also affect *HBG2* gene

25
Zhong, H., Yang, X., Kaplan, L. M., Molony, C. & Schadt, E. E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.* 86, 581–91 (2010).
26
Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–13 (2007).
27
Silverberg, M. S. *et al.* Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.* 41, 216–20 (2009).
28
Abraham, C. & Cho, J. H. Inflammatory bowel disease. *N. Engl. J. Med.* 361, 2066–78 (2009).
29
Huynh, D. *et al.* Colony stimulating factor-1 dependence of paneth cell development in the mouse small intestine. *Gastroenterology* 137, 136–44, 144.e1–3 (2009).
30
Mombaerts, P. *et al.* Spontaneous development of inflammatory bowel disease in T cell receptor mutant mice. *Cell* 75, 274–82 (1993).
31
Eike, M. C. *et al.* Genetic variants of the HLA-A, HLA-B and AIF1 loci show independent associations with type 1 diabetes in Norwegian families. *Genes Immun.* 10, 141–50 (2009).
32
Valdes, A. M. & Thomson, G. Several loci in the HLA class III region are associated with T1D risk after adjusting for DRB1-DQB1. *Diabetes. Obes. Metab.* 11 Suppl 1, 46–52 (2009).
33
Yang, B., Houlberg, K., Millward, A. & Demaine, A. Polymorphisms of chemokine and chemokine receptor genes in Type 1 diabetes mellitus and its complications. *Cytokine* 26, 114–21 (2004).

expression levels in *trans* (at FDR 0.05), while other pairs of MCV SNPs converge on *ESPN, VWCE, PDZK1IP1* and *RAP1GAP*.

For mean platelet volume (MPV) we observed that MPV SNPs rs12485738 on 3p26 and rs11602954 on 11p15 affect several transcripts in *trans*. These two SNPs converge on *GP9, F13A1, C19orf33, SAMD14, VCL* and *GNG11*. As *GP9* and *F13A1* are known blood coagulation genes, *C19orf33* is a potential candidate gene, involved in coagulation as well. This is substantiated by strong co-expression between *GP9* and *C19orf33* within peripheral blood (Pearson r = 0.45, p = 7.0 x 10^{-63}) and the fact these SNPs independently also affect various other blood coagulation genes in trans (including *CD151, GP1BB, ITGA2B, MMRN1, THBS1* and *VWF*, Figure 4). Many of these are specific to megakaryocytes that are platelet precursor cells[43]. As expected, the Gene Ontology term 'blood coagulation' is strongly overrepresented among all these *trans*-genes, Fisher's exact p = 1.0 x 10^{-10}.We observed that MPV SNP rs12485738 (on 3p14.3) was also *trans*-acting on tropomyosin 1 (*TPM1*, 15q22.2, p = 9.7 x 10^{-9}), a gene that is also regulated in *cis* by another MPV variant (rs11071720 on 15q22.2, p = 1.4 x 10^{-13}). We observed this for two different expression probes that map within different locations of the *TPM1* transcript (probes 5560246 and 610519), and note strong co-expression for these two *TPM1* probes with 46 MPV *trans*-genes (Pearson r > 0.19, p < 1.0 x 10^{-11}, including five known coagulation genes). Although several genes reside within the rs11071720 MPV locus, these observations strongly implicate *TPM1* as the causal MPV gene. For both MPV and MCV we observed that the identified *cis*- and *trans*-eQTL probes

34
Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447, 1087–93 (2007).

35
Van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–6 (2002).

36
Yu, J. X. *et al.* Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer* 7, 182 (2007).

37
Gilman, J. G. & Huisman, T. H. DNA sequence variation associated with elevated fetal G gamma globin production. *Blood* 66, 783–7 (1985).

38
Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* 39, 1197–9 (2007).

39
Thein, S. L. *et al.* Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl. Acad. Sci. U. S. A.* 104, 11346–51 (2007).

40
Nuinoon, M. *et al.* A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. *Hum. Genet.* 127, 303–14 (2010).

41
Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* 41, 1191–8 (2009).

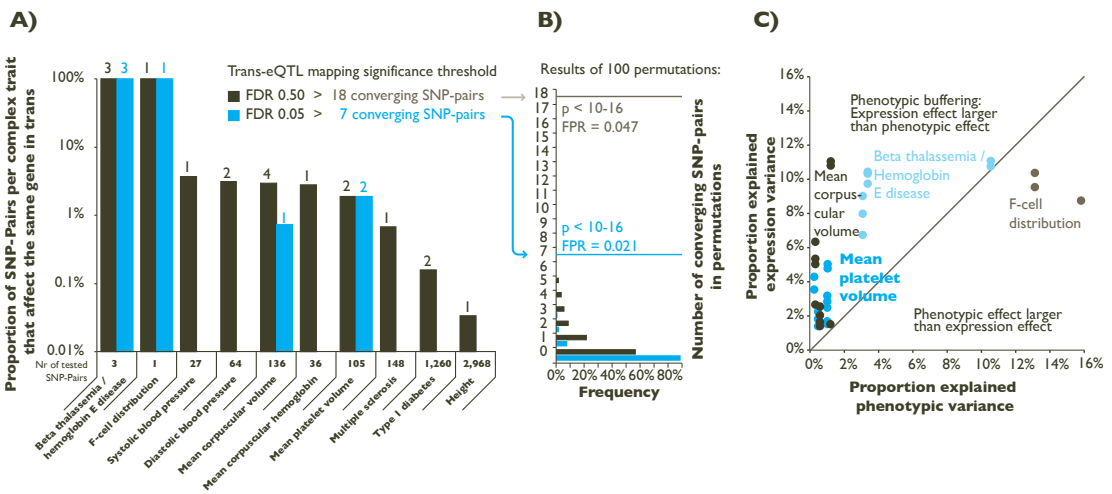


Figure 4. Pairs of SNPs that cause the same phenotype more frequently than expected also affect the same downstream genes.

Various pairs of unlinked SNPs cause the same phenotype but also converge on the same downstream genes.

A) When using *cis*- and *trans*-eQTLs, identified when controlling FDR at 0.05, 7 unique pairs of SNPs cause the same phenotype but also affect the same downstream gene. When controlling the FDR at 0.50 for the *trans*-eQTLs, 18 unique pairs of SNPs show this convergence. B) This is significantly higher than expected, determined using 100 permutations.

C) The SNPs that affect these downstream genes in most instances explain a proportion of the downstream gene expression variation that is substantially higher than what their effect is on the eventual phenotypes.

generally were more strongly co-expressed in peripheral blood than expected (Figure S4, MPV co-expression Wilcoxon $P < 10^{-200}$, MCV co-expression Wilcoxon P = 0.009) substantiating the likelihood these genes reflect coherent biological sets.
We repeated this co-expression analysis after we had regressed out all *cis*- and *trans*-eQTL effects, and observed that most of this co-expression was independent of the eQTL SNP-effect on the expression of these genes, which further substantiates that these genes are biologically related (MPV co-expression Wilcoxon $P < 1^{-200}$, MCV co-expression Wilcoxon P = 0.018).

Phenotypic buffering

Although the observed convergence provides insight into downstream genes, it is not clear whether the MPV or MCV phenotypes are eventually caused through these *trans*-genes, or whether these *trans*-eQTLs emerged as a result of changes to the volume of the platelets or the erythrocytes.

In order to gain insight into this, we analyzed the effect size of these SNP variants on both the expression levels and the phenotypes. While the effect sizes of these trait-associated SNPs on eventual phenotypes were usually small, their intermediate (molecular) effect was often substantially larger. This supports the notion that the effect on e.g. MPV and MCV is through these *trans*-genes, and suggests the presence of 'phenotypic buffering', shown previously in plants[44], in humans (Table 2, Figure 4B): the effects of the 18 converging pairs of SNPs on gene expression levels were often substantially higher than the originally reported effect sizes on the trait-phenotypes. For example, several MPV-and MCV-associated SNPs explain between 1.41% and 10.99% of *trans*-expression variation within the 1,469 unrelated samples, whereas these SNPs only explain between 0.24% and 1.12% of the MPV and MCV phenotype variation (and as such required over 13,000 samples[41,42] for identification, Figure 4B).

Replication of *trans*-eQTLs in monocytes and four additional primary tissues

We analyzed peripheral blood which is a mixture of different hematopoietic cell types. In addition, we also assessed whether the identified trait-associated *trans*-eQTLs (detected at FDR 0.05) could be replicated in a single cell-type dataset. This is an important question, as it is potentially possible that the *trans*-acting SNP are able to alter the amount, volume or ratio of certain blood cell types, which might as a consequence result in an indirect net effect on the measured gene expression levels within the mix of the cells that comprise whole blood.

We therefore analyzed monocyte expression data from 1,490 independent samples[45] and did not find evidence that this was a widespread phenomenon as we could replicate 46 out of the 130 different *trans*-eQTLs (each of these with a nominal $p < 1.0 \times 10^{-5}$ in the monocyte data, Table S8). These replicated eQTLs include the genes *AOAH, HBG2, GP9, F13A1, SAMD14, CD151, ITGA2B, MMRN1, THBS1, VWF* and *TPM1* mentioned above. Surprisingly we could also replicate the *trans*-eQTL effects on various blood-coagulation genes for mean platelet volume SNP rs12485738: one might argue that rs12485738 primarily increases platelet volume, resulting in a relatively higher volume of platelet-RNA

42
Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* 41, 1182–90 (2009).
43
Watkins, N.A. *et al.* A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood* 113, e1–9 (2009).
44
Fu, J. *et al.* System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat. Genet.* 41, 166–7 (2009).
45
Zeller, T. *et al.* Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One* 5, e10693 (2010).

Table 2.

Trait-associated SNPs converge on the same downstream genes.

Complex trait	Unlinked SNP-pair		Explained trait variance	
	SNP1	SNP2	SNP1	SNP2
ß thalassemia	rs766432	rs2071348	3.30%[40]	3.00%[40]
	rs766432	rs2071348	3.30%[40]	3.00%[40]
	rs766432	rs2071348	3.30%[40]	3.00%[40]
	rs9376092	rs766432	10.50%[40]	3.30%[40]
	rs9376092	rs2071348	10.50%[40]	3.00%[40]
	rs9376092	rs766432	10.50%[40]	3.30%[40]
	rs9376092	rs2071348	10.50%[40]	3.00%[40]
F-cell distribution	rs1427407	rs9399137	13.10%[38]	15.80%[38]
	rs1427407	rs9399137	13.10%[38]	15.80%[38]
Mean corpuscular volume	rs12718597	rs643381	0.26%[41]	0.50%[41]
	rs2540917	rs643381	0.24%[41]	0.50%[41]
	rs4895441	rs2540917	1.12%[41]	0.24%[41]
	rs4895441	rs2540917	1.12%[41]	0.24%[41]
	rs4895441	rs643381	1.12%[41]	0.50%[41]
	rs643381	rs4895441	0.50%[41]	1.12%[41]
Mean corpuscular Hb	rs628751	rs7776054	0.34%[41]	1.02%[41]
Mean platelet volume	rs12485738	rs11602954	0.93%[42]	0.41%[42]
	rs12485738	rs11602954	0.93%[42]	0.41%[42]
	rs12485738	rs11602954	0.93%[42]	0.41%[42]
	rs12485738	rs11071720	0.93%[42]	0.18%[42]
	rs12485738	rs11071720	0.93%[42]	0.18%[42]
	rs12485738	rs11602954	0.93%[42]	0.41%[42]
	rs12485738	rs11602954	0.93%[42]	0.41%[42]
	rs12485738	rs11602954	0.93%[42]	0.41%[42]
Height	rs910316	rs10946808	N/A	N/A
Multiple sclerosis	rs2523393	rs9271366	N/A	N/A
Systolic blood pressure	rs3184504	rs2681492	N/A	N/A
Diastolic blood pressure	rs3184504	rs2681472	N/A	N/A
	rs653178	rs2681472	N/A	N/A
Type 1 diabetes	rs9272346	rs11171739	N/A	N/A
	rs9272346	rs1701704	N/A	N/A

Indicated are 18 pairs of unlinked SNPs that are associated with the same complex phenotype and that also affect the expression levels of the same downstream gene(s) in *cis* (FDR 0.05) or *trans* (FDR 0.50). [a] Erythrocyte specific gene according to HaemAtlas[43]. [b] Megakaryocyte specific gene according to HaemAtlas[43]. [c] Explained phenotypic variation is shown for traits when reported in the original papers (indicated in superscript) that describe these SNP – phenotype association.

SNP-pair convergences		eQTL significance		Explained expression variance	
SNP1	SNP2	SNP1	SNP2	SNP1	SNP2
HBG2	4010040	2.12×10^{-29}	4.22×10^{-24}	9,72%	7,96%
HBG2[a]	450537	7.67×10^{-37}	6.46×10^{-24}	10,41%	6,72%
HBG2	6400079	6.95×10^{-07}	3.80×10^{-06}	10,28%	8,99%
HBG2	4010040	1.73×10^{-32}	2.12×10^{-29}	10,75%	9,72%
HBG2	4010040	1.73×10^{-32}	4.22×10^{-24}	10,75%	7,96%
HBG2[a]	450537	9.49×10^{-39}	7.67×10^{-37}	11,06%	10,41%
HBG2[a]	450537	9.49×10^{-39}	6.46×10^{-24}	11,06%	6,72%
HBG2	4010040	1.18×10^{-28}	1.70×10^{-26}	9,51%	8,74%
HBG2[a]	450537	1.21×10^{-36}	1.86×10^{-30}	10,35%	8,75%
VWCE	1450608	3.39×10^{-10}	1.74×10^{-06}	2,65%	1,61%
ESPN	3440630	1.95×10^{-15}	6.20×10^{-07}	4,99%	1,99%
HBG2	4010040	2.74×10^{-32}	2.87×10^{-19}	10,71%	6,32%
HBG2[a]	450537	1.31×10^{-38}	3.19×10^{-18}	10,99%	5,29%
RAP1GAP[a]	4890181	2.46×10^{-06}	5.57×10^{-06}	1,51%	1,41%
PDZK1IP1	3170270	7.44×10^{-10}	4.27×10^{-06}	2,55%	1,45%
PDZK1IP1	3170270	7.74×10^{-10}	8.97×10^{-07}	2,55%	1,65%
GP9[b]	1050292	3.62×10^{-17}	1.14×10^{-07}	4,82%	1,93%
GNG11	1580025	9.67×10^{-12}	2.23×10^{-06}	3,22%	1,52%
F13A1	2230241	5.37×10^{-09}	3.13×10^{-09}	2,54%	2,38%
TPM1	5560246	1.47×10^{-08}	1.38×10^{-13}	2,58%	4,32%
TPM1	610519	1.45×10^{-06}	4.41×10^{-13}	1,60%	3,58%
SAMD14[b]	5560280	4.08×10^{-18}	3.10×10^{-06}	5,05%	1,47%
C19orf33	630470	6.26×10^{-11}	1.16×10^{-08}	2,86%	2,37%
VCL[b]	70592	7.49×10^{-07}	6.81×10^{-06}	1,72%	1,39%
BTN3A2	4610674	5.42×10^{-06}	9.79×10^{-10}	1,40%	2,60%
TGFBR2	2340324	5.15×10^{-07}	1.07×10^{-06}	2,01%	1,90%
LOC338758	6650035	1.28×10^{-06}	9.17×10^{-08}	1,87%	2,27%
LOC338758	6650035	1.28×10^{-06}	2.23×10^{-08}	1,87%	2,49%
LOC338758	6650035	1.54×10^{-06}	2.23×10^{-08}	1,85%	2,49%
KRT18	6580270	1.87×10^{-06}	4.72×10^{-06}	2,06%	1,70%
KRT18	6580270	1.87×10^{-06}	9.40×10^{-06}	2,06%	1,39%

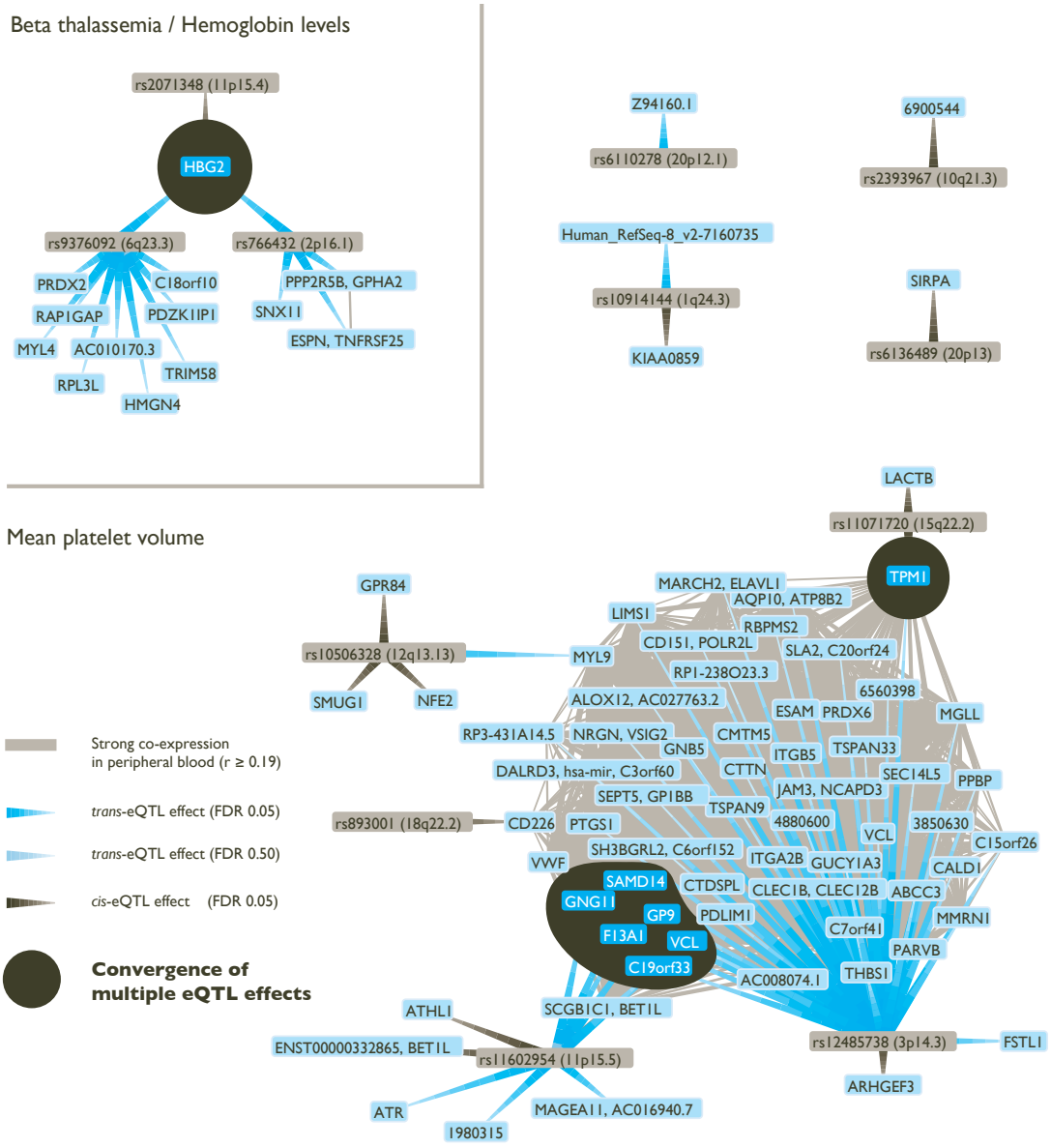
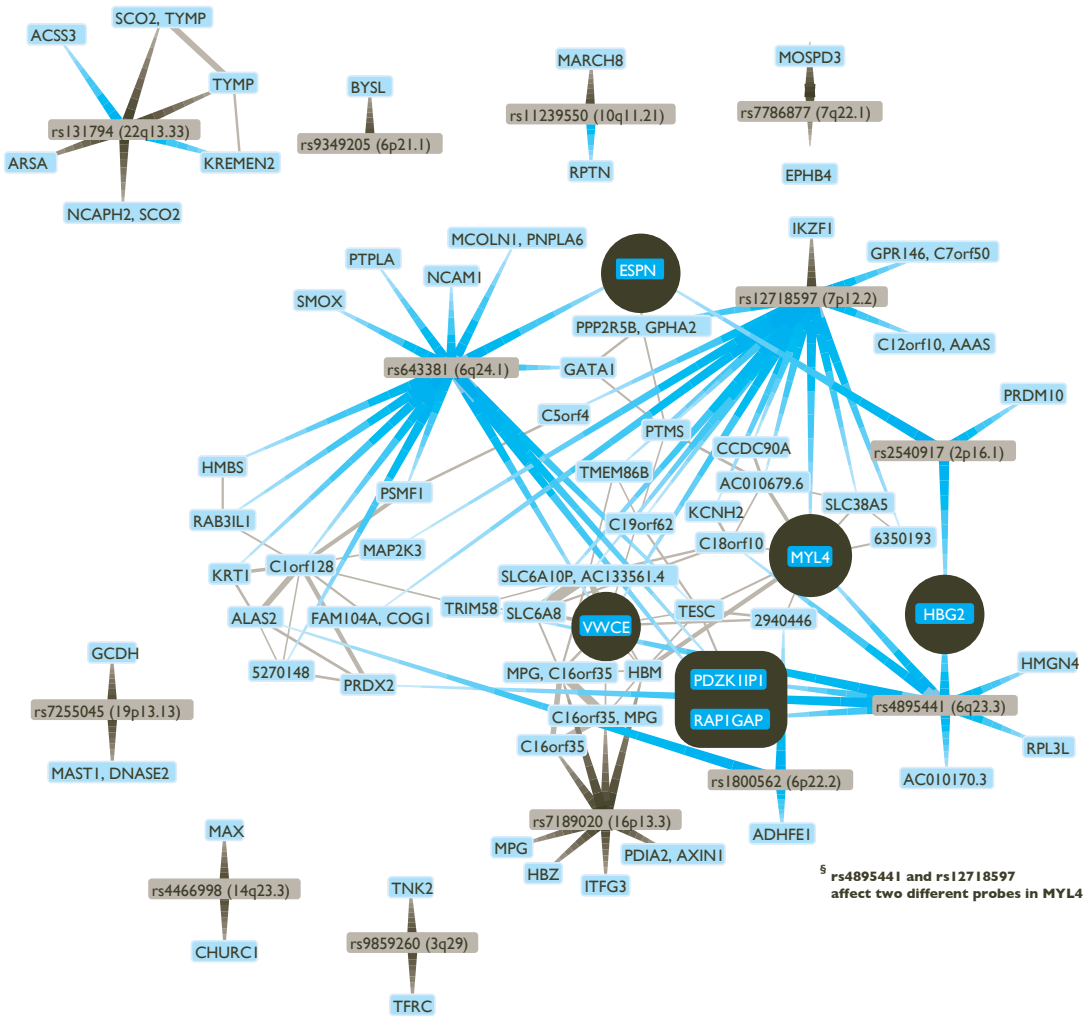


Figure 5. Trait-associated SNPs show convergence on multiple genes.

For several traits different and unlinked SNPs affect the same genes in *cis* or *trans*. For beta thalassemia three different loci affect hemoglobin (*HBG2*) gene expression (one in *cis*, indicated with gray arrow, two in *trans* (at FDR 0.05), indicated with cyan arrows). For mean corpuscular volume (MCV) the same *trans*-effects on *HBG2* (at FDR 0.05) exist, but convergence is also apparent on *ESPN, VWCE, PDKZ1IP1* and *RAP1GAP* (at FDR 0.50). For mean platelet volume (MPV) numerous *trans*-effects on genes involved in blood coagulation were identified. Two MPV loci (rs12485738 on 3p26.3 and rs11602954 on 11p15.5) both affect *GP9, F13A1* and *C19orf33* (at FDR 0.05) and *SAMD14, GNG11* and *VCL* (at FDR 0.50). Peripheral blood co-expression (Pearson correlation coefficient r≥0.19, $p < 1.0 \times 10^{-11}$) between genes is indicated in light grey.

Mean corpuscular volume



§ **rs4895441 and rs12718597 affect two different probes in MYL4**

when assessing total peripheral blood RNA. If this were to be the case, a measurable *trans*-effect is expected for platelet-specific (blood coagulation) genes in whole blood. Such an effect would then not actually be an expression-QTL, but rather a 'cellular-QTL'. However, the *trans*-eQTLs for rs12485738 were also present in single cell-type monocyte datasets, indicating that the above concerns do not apply. Clearly, *trans*-eQTL effects can manifest themselves outside the primary cell-type, in which they are expected to operate.

We also replicated 18 trait-associated *trans*-eQTLs (including *AOAH*, detected at FDR 0.05) in an independent dataset comprising four different non-blood tissues (subcutaneous adipose, visceral adipose, liver and muscle, Figure S5, Table S9 and S10). Since this dataset comprised only 90 samples, it is very encouraging that 18 *trans*-eQTL could be replicated.

Discussion

Here we investigated gene expression in peripheral blood from 1,469 individuals to identify *cis*- and *trans*-effects of common variants on gene expression levels. When comparing to other genetical genomics studies[12–14,16,18,21–24,45,46] we observe an increasing percentage of genes that are *cis*- or *trans*-regulated (39% of 19,689 unique genes at FDR 0.05). When eQTL studies further increase the sample-sizes and thus increase statistical power, we expect that for the far majority of genes the expression levels are to some extent determined by genetic variation.

GWA studies have identified many loci, but it is still often unclear what the affected gene in each locus is. Here we showed that 39% of trait-associated SNPs affect gene expression in *cis* which is helpful in pinpointing the most likely gene per susceptibility locus. However, GWAS do not immediately provide insight in the *trans*-effects of these susceptibility variants on downstream genes. Here we identified for 2.6% of all trait-associated SNPs *trans*-eQTL effects on in total 113 unique genes. While some of these *trans*-eQTLs are known to be involved in these phenotypes (such as *HBG2* in hemoglobin protein levels and β-Thallasemia), most of these genes have not been implicated before in these complex traits, and provide additional insight in the downstream mechanisms of these variants. Interestingly, 48% of *trans*-acting trait-associated SNPs map within the HLA, indicating the HLA has a prominent role in regulating peripheral blood gene expression. This might partly explain why the HLA has been found to be associated with so many different diseases.

While we concentrated on peripheral blood, we could replicate 35% of the *trans*-eQTLs in monocytes. Particularly surprising was the observation that for SNPs, known to affect the volume of platelets or erythrocytes the identified *trans*-eQTL effects in whole blood were also present in these monocytes. Among these replicated genes are a considerable number of highly plausible *trans*-genes. For example, for mean platelet volume SNP rs12485738 we detected the same *trans*-eQTL effects on seven well-known blood coagulation genes (*F13A1, GP1BB, GP9, ITGA2B, MMRN1, THBS1* and *VWF*) in both the peripheral blood data and

46
Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246–50 (2009).

the monocyte data. Interestingly, in both datasets, *trans*-effects for this SNP on another 31 genes were identified as well, which suggests these genes play a role in blood coagulation. It can thus be concluded that *trans*-eQTLs, identified in peripheral blood, generally apply to monocytes as well. We assumed these eQTLs might therefore also be present in other, non-blood tissues, as previously observed for rodents[47–49]. Indeed we could replicate some of these *trans*-eQTLs in a smaller dataset of four non-blood tissues. Importantly, as mentioned before[46], the allelic directions were nearly always identical to blood, which implies that *trans*-eQTLs, if also present in another tissue, work in the same way.

Our observation that sets of independent SNPs, associated with the same complex phenotype sometimes also affect exactly the same *trans*-gene, further substantiates the validity of our findings. Based on the reported effect-sizes of these variants on these complex phenotypes, we have shown here that the individual effects of these SNPs on *trans*-gene expression can often be stronger. This suggests that these down-stream gene expression effects do not fully propagate to the eventual phenotype and are somehow buffered. This 'phenotypic buffering' has been observed before in plants[44] and suggests that additional compensatory mechanisms exist that control these complex phenotypes. However, we do realize that accurate estimates on this phenomenon requires the availability of both gene-expression and phenotype data for these traits. As we did not have these phenotypes for our samples, we relied upon estimates from literature. Future studies that have collected both genome-wide genotype, expression and phenotype data from the same individuals will permit answering the question what the extent of this phenotypic buffering is. We should emphasize that the number of converging pairs of SNPs that we identified must be a very strong underestimate, and as such the false-negative rate from this analysis is likely to be high: As we observed that on average 40.4% of the trait-associated SNPs affect gene expression levels in *cis*, we expect that many of these SNPs will exert effects on gene expression in *trans*. However, these effects are likely to be small and due to multiple testing issues our current study identified only a relatively small set of *trans*-eQTL effects. Likewise the number of detected converging pairs of SNPs is even smaller. However, as we observed this convergence for various pairs of SNPs, future genetical genomics studies using larger sample sizes will likely reveal many more pairs of converging SNPs, providing better insight in the downstream molecular mechanisms that are affected by these disorders.

The convergence and phenotypic buffering we observed might also help uncover some of the missing heritability in complex disease. As there are probably many SNPs with low marginal phenotypic effects[50], GWAS currently lack power to detect these. However, the effect of these trait-associated SNPs on expression seems to be less multifactorial, leading to larger expression effects. These numerous expression disturbances will eventually converge to a phenotype, explaining the small phenotypic effect of individual trait-associated SNPs.

47
Peirce, J. L. *et al.* How replicable are mRNA expression QTL? *Mamm. Genome* 17, 643–56 (2006).

48
Petretto, E. *et al.* New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol.* 6, e1000737 (2010).

49
Petretto, E. *et al.* Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet.* 2, e172 (2006).

50
Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–9 (2010).

Therefore, studying expression as intermediate phenotype will be important for disease association studies trying to account for the missing heritability of complex diseases. Disease SNPs, already found to be disease-associated and marked as eQTL, lead to a set of candidate downstream genes. Additional genetic variants that also affect the expression of these genes will therefore be powerful candidates for disease susceptibility.

Materials and Methods

Peripheral blood genetical genomics study populations

The peripheral blood genetical genomics study population contained 1,469 unrelated individuals from the United Kingdom and the Netherlands. Some of these are healthy controls while others are patient samples. The 49 ulcerative colitis (UC) cases in this study are part of the inflammatory bowel disease (IBD) cohort of the University Medical Centre Groningen. The 111 celiac disease samples were collected within the Barts and the London NHS Trust and the Oxford Radcliffe Hospitals NHS Trust. The 453 chronic obstructive pulmonary disease (COPD) samples were collected within the NELSON study. The 856 amyotrophic lateral sclerosis (ALS) cases and controls were collected in the University Medical Centre Utrecht. All samples were collected after informed consent and approved by local ethical review boards. Individual sample information is provided in Table S11.

Peripheral blood (2.5 ml) for all samples was collected with the PAXgene system (PreAnalytix GmbH, UK). PAXgene vials were chosen to prevent density gradient centrifugation, immortalization or in vitro cell culture artifacts changing mRNA profiles. PAXgene tubes were mixed gently and incubated at room temperature for two hours. After collection, tubes were frozen at 220°C for at least 24 hours followed by storage at 280°C. RNA was isolated using the PAXgene Blood RNA isolation kit (PreAnalytix GmbH, UK). RNA was quantified using the Nanodrop (Nanodrop Technolo-gies, USA). Total RNA integrity was analyzed using an Agilent Bioanalyzer (Agilent Technologies, USA).

Peripheral blood SNP genotyping

Peripheral blood samples were either genotyped using the Illumina (Illumina, San Diego, USA) HumanHap300, Human-Hap370 or 610 Quad platform. Genotyping was performed according to standard protocols from Illumina. Although the different genotype oligonucleotide arrays differ, they share 294,757 SNPs, to which the analysis was confined. In addition, SNPs with a minor allele frequency of < 5%, or a call-rate < 95%, or deviating from Hardy-Weinberg equilibrium (exact p-value < 0.001) were excluded, resulting in 289,044 SNPs for further analysis. Genotype calling for each SNP was performed by a previously described algorithm[51].

Peripheral blood Illumina expression profiling

Anti-sense RNA was synthesized, amplified and purified using the Ambion Illumina TotalPrep Amplification Kit (Ambion, USA) following the manufacturers' protocol. Complementary RNA was either hybridized to Illumina HumanRef-8 v2 arrays (229 samples, further referred to as H8v2) or Illumina HumanHT-12 arrays (1,240

51 Van Heel, D.A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* 39, 827–9 (2007).

samples, further referred to as HT12) and scanned on the Illumina BeadArray Reader. Raw probe intensities were extracted using Illumina's BeadStudio Gene Expression module v3.2 (No background correction was applied, nor did we remove probes with low expression). The raw expression data of the 1,240 HT12 peripheral blood samples were combined with the raw expression data of 296 replication samples (described in detail in paragraph '*Trans*-eQTL replication dataset'). Both datasets (H8v2 and HT12) were quantile normalized separately to the median distribution and expression values were subsequently $\log_2$ transformed. Subsequently, the probes were centered to zero and linearly scaled such that each probe had a standard deviation of one.

Integration of the Illumina H8V2 and HT12 peripheral blood expression platform identifiers

The HT12 and H8v2 arrays share a considerable number of probes with identical probe sequences. However, in a considerable number of occasions the two platforms use different probe identifiers for the same probe sequences. More importantly, although probe identifiers are often identical, they sometimes represent different probe sequences. In order to permit a meta-analysis incorporating data from both arrays, we decided on the following naming convention: if an H8v2 probe had the same sequence as an HT12 probe, the HT12 'ArrayAddressID' probe identifier was used. If not, the original H8v2 probe identifier was used, but with the prefix "Human_RefSeq-8_v2-" to prevent any potential probe identifier ambiguity. A total of 52,061 unique probes were used for further analysis, representing 19,609 unique genes according to HUGO gene nomenclature.

Initial genomic mapping of Illumina expression probe sequences

Various mapping strategies were used for the expression probes to get a mapping location that was as unambiguous as possible: if probes have been mapped incorrectly, or cross-hybridize to multiple genomic loci, it might be that an eQTL will be incorrectly deemed a *trans*-eQTL, while in fact it is a *cis*-eQTL or primer polymorphisms. We used Ensembl database version 52 (NCBI 36.3 assembly) to obtain, for each annotated gene, the transcript with the largest number of exons and included this main spliced transcript in our reference set. Second, we added one sequence per intron, extending intron boundaries 40 bp on each side to allow mapping of the 50 bp probe sequences that overlapping exon-intron junctions. Last, a version of the reference DNA genome with masked annotated transcripts was included. Probe sequences were mapped using NOVOALIGN V2.05.12 for all the sequences (main transcript, introns, and non-standard exon-exon junctions) originating from the same transcript (parameters 2t 150 2v 20 20 200 [.]([^_]*)_). For each probe it was determined whether it was mapping uniquely to one particular genomic locus, or, if multiple hits were present whether all these mappings resided in each other vicinity (< 250 kb). Probes that did not map at all, or mapped to multiple different loci were excluded from further analyses. Using this approach, 43,202 of the 48,751 probes on the HT12 and 21,316 of the 22,185 probes on the H8v2 platform were eventually mapped to a single genomic location.

eQTL mapping

In order to detect *cis*-eQTLs, analysis was confined to those probe-SNP combinations for which the distance from the probe transcript midpoint to SNP genomic location was ≤ 250 kb. For *trans*-eQTLs, analysis was confined to those probe-SNP combinations for which the distance from probe transcript midpoint to SNP genomic location was ≥ 5 Mb (to exclude the possibility of accidentally detecting *cis*-eQTLs due to long ranging linkage disequilibrium). Additionally, for the *trans*-eQTL analysis the effects of the significant *cis*-eQTLs were removed from the expression data by keeping the residual expression after linear regression.

Association for *cis*- and *trans*-eQTL was tested with a non-parametric Spearman's rank correlation. For directly genotyped SNPs we coded genotypes as 0, 1 or 2, while for imputed SNPs we used SNP dosage values, ranging between 0 and 2. When a particular probe-SNP pair was present in both the HT12 and H8v2 datasets, an overall, joint p-value was calculated using a weighted (square root of the dataset sample number) Z-method.

To correct for multiple testing, we controlled the false-discovery rate (FDR) at 0.05: the distribution of observed p-values was used to calculate the FDR, by comparison with the distribution obtained from permuting expression phenotypes relative to genotypes 100 times within the HT12 and H8v2 dataset for both the *cis*- and *trans*-analyses[52].

In order to increase the number of detectable *cis*- and *trans*-eQTLs we applied a principal component analysis (PCA) on the sample correlation matrix. We, among others[19,20], argue that the dominant PCs, capturing the larger part of the total variation, will primarily capture sample differences in expression that reflect physiological or environmental variation as well as systematic experimental variation (e.g. batch and technical effects). Figure S6 shows for the 1,240 HT12 samples what per individual the PC scores are. It is evident there are, especially among the first PCs, strong batch effects are still present after proper quantile-quantile normalization. By removing the variation captured by these PCs, we expected that the residual expression is more strongly determined by genetic variants and the number of significantly detected *cis*- and *trans*-eQTLs will increase. An aspect to consider is that with the removal of more PCs from the data, the degrees of freedom of the data will decrease. Furthermore, it is not immediately clear which PCs will actually capture physiological, environmental, and systematic variation, which might lead to removal of genetically determined expression variation as well. Therefore a tradeoff has to be made on the number of PCs to subtract from the data. We assessed this systematically, by removing up to 100 PCs from the genetical genomics dataset (in steps of 5).

Figure S7A shows that the number of significantly detected *cis*-eQTL probes increases two-fold when 50 PCs were removed from the expression data. There is a long plateau visible (around PC50), where the number of detected *cis*-eQTLs probes remains approximately constant, irrespective of removing for instance 10 fewer or 10 extra PCs (reported numbers in this figure also include false-positive eQTLs due to potential primer polymorphisms, as we

52 Breitling, R. *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 4, e1000232 (2008).

here wanted to solely compare the performance of removing different numbers of PCs). Figure S7B shows that of the initial 5,950 significantly detected *cis*-eQTL probes (no PCs removed), 4,965 (83.5%) were still detected with 50 PCs subtracted. The 985 initially detected *cis*-eQTLs probes, yet no longer detected when 50 PCs had been removed from the expression data, all had a low significance (Figure S8). As we controlled the FDR at 0.05 in all analyses it is therefore likely that a considerable amount of these reflect false-positives. Figure S8C shows that for all the overlapping 4,965 detected *cis*-eQTLs probes between the different analyses, the allelic direction was identical, and effect size on expression correlate well (Pearson r = 0.95) although these were nearly always stronger after having subtracted 50 PCs.

We assessed this for *trans*-eQTLs as well. An important aspect to consider is that *trans*-eQTL SNPs might affect multiple genes. If these effects are substantial (either in effect size or the number of affected genes), it is likely that a certain PC will capture this. Removal of such PCs from the expression data will therefore unintentionally result in the inability to detect these *trans*-eQTLs. In order to avoid such false-negatives we first performed a QTL analysis on the first 50 PCs (that had been removed from the expression data for the *cis*-eQTL analysis) to assess whether some of these PCs are under genetic control (genome-wide analysis, controlling FDR at 0.05). We did this for the large HT12 and the smaller H8v2 expression data separately, as PCA had been applied independently to these datasets. We observed that out of the first PCs in the HT12 data three PCs and in the H8v2 two PCs were to some extent genetically determined (r2 > 5%). This was different for PCAs 26–50 in the HT12 data: 11 PCs were under substantial genetic control (Figure S9a). We therefore assumed that most *trans*-eQTLs could be detected when removing approximately 25 PCs. We quantified this systematically, by removing increasing amounts of PCs from the expression data and conducting a full genome-wide *trans*-eQTL mapping. Indeed, in these analyses at most 244 significant *trans*-eQTLs could be detected (at FDR 0.05, with potential false-positives due to cross-hybridizations removed), when removing 25 PCs (Figure S9b). The overlap with the expression with no PCs removed was substantial: 62 of the 82 *trans*-eQTLs (77%), detected in the original analysis were detected as well in the analysis with 25 PCs removed (Figure S9c), all with identical allelic directions (Figure S9d).

Identification of false eQTLs due to primer polymorphisms and cross-hybridization

One should be aware that sequence polymorphisms can cause many false *cis*-eQTLs[53]. Such false *cis*-eQTLs do not reflect actual expression differences caused by sequence polymorphisms in *cis*-acting factors that affect mRNA levels. Instead they indicate hybridization differences caused by sequence polymorphisms in the mRNA region that is targeted by the microarray expression probes. Therefore, SNP-probe combinations were excluded from the *cis*-eQTL analysis when the 50 bp long expression probe mapped to a genomic location that contained a known SNP that was showing at least some LD (r2 > 0.1) with the *cis*-SNP. We used SNP data from the 1000 Genomes Projects, as it contains LD information for 9,633,115 SNPs (April 2009 release, based on 57 CEU samples of European descent).

53 Alberts, R. *et al.* Sequence polymorphisms cause many false cis-eQTLs. *PLoS One* 2, e622 (2007).

Detected *trans*-eQTLs might also reflect false-positives, although we initially had attempted to map the expression probes as accurately as possible, by using the aforementioned three different mapping strategies: it is still well possible that some of the identified, putative *trans*-eQTLs in fact reflect very subtle cross-hybridization (e.g. pertaining to only a small subsequence of the probe). We therefore tried to falsify each of the putative *trans*-eQTLs by attempting to map each *trans*-probe into the vicinity of the SNP probe location, by using a highly relaxed mapping approach. All putative Illumina *trans*-expression probes were mapped using SHRiMP[54], which uses a global alignment approach, to the human reference genome (NCBI 36.3 build). The mapping settings were chosen very loosely to permit the identification of nearly all potential hybridization locations: match score was 10, the mismatch score was 0, the gap open penalty was 2250, the gap extension penalty was 2100, Smith and Waterman minimum identical alignment threshold was 30.0%, while other SHRiMP parameters were left at default. Using these settings all mappings with a minimum overlap of 15 bases, or with 20 matches with one mismatch, or 30 matches with 2 mismatches, or full-length (50 bp) probe hybridizations with no more than 15 mismatches were accepted. Any *trans*-eQTL was discarded, if the expression probe had a mapping that was within 2 Mb of the SNP that showed the *trans*-eQTL effect. Once these potential false-positive *trans*-eQTLs had been removed from the real, non-permuted data, we repeated the multiple testing correction (again controlling the FDR at 0.05).

Using this strategy we observed several instances where only 20 out the 50 bases of a probe sequence mapped in the vicinity of the *trans*-SNP (data not shown). For these *trans*-eQTLs the Spearman's rank correlation p was often lower than 10-100, which would imply these SNPs explain over 25% of the total expression variation of the corresponding *trans*-genes. Given the small amount of *trans*-eQTLs we detected in total, such effect sizes are quite unlikely and therefore provide circumstantial evidence these indeed reflect cross-hybridization artifacts.

We also assessed whether any of the Illumina SNPs that constitute *trans*-eQTLs might map to a different position than what is reported in dbSNP. As such we mapped the 50 bp Illumina SNP probe sequences to the genome assembly, permitting up to four mismatches per 50 bp SNP probe sequence. We did not observe any SNP that could map (with some mismatches) to the same chromosome of the *trans*-probe.

It is still possible that some of the *trans*-eQTLs for which we did not find any evidence of cross-hybridization, still are false positives, by missing some cross-hybridizations due to imperfections in the NCBI v36 assembly we used. Although we have identified numerous occasions where a SNP affects two different probes within the same gene in *trans*, substantiating the likelihood these *trans*-eQTLs are real, providing unequivocal evidence that all our reported *trans*-eQTLs are real is not straightforward.

[54] Rumble, S. M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* 5, e1000386 (2009).

Enrichment analysis of trait-associated SNPs and SNPs located within the HLA region

To assess enrichment of trait-associated SNPs, we used a collection of 1,262 unique SNPs from 'A Catalog of Published Genome-Wide Association Studies' (accessed 09 February 2010, and each having at least one reported association p-value < 5.0×10^{-7}). We could successfully impute the genotypes for 1,167 of these SNPs and therefore confined all analyses to these SNPs. Of these SNPs 572 had been directly genotyped on the Illumina HumanHap300 platform, with a MAF < 0.05, an HWE exact p-value < 0.0001 and call-rate < 95%.

To ascertain whether these SNPs are more often constituting an eQTL than expected, we used a methodology that is not affected by the following potential confounders: non-even distribution of SNP markers and expression probe markers across the genome, differences in MAF between SNPs and LD structure within the genotype date and correlation between probes in the expression data. Additionally, this methodology is also not confounded by the fact that for certain traits different SNPs in strong LD can have been reported, due to differences in the platforms that were used to identify these loci. We first determined how many unique eQTL SNPs had been identified in the original eQTL mapping (with an FDR < 0.05) and how many of these are trait-associated. Subsequently we permuted the expression phenotypes relative to the genotypes (thus keeping the correlation structure within the genotype data and the correlation structure within the expression data intact, yet assigning the genotypes of a sample to the expression data of a randomly chosen sample) and reran the eQTL mapping, sorting all tested eQTLs on highest significance. We then took an equal number of top associated, but permuted, eQTL SNPs and determined how many of these permuted eQTL SNPs are trait-associated. By performing 100 permutations we obtained an empiric distribution of the number of trait-associated SNPs expected by chance.
We subsequently fitted a generalized extreme value distribution (EVD, using the EVD add-on package for R), permitting us to estimate realistic enrichment significance estimates (called EVD p throughout the manuscript).

For the MHC enrichment analysis the followed procedure was identical, with the difference that we looked for enrichment for SNPs within the MHC, defined as SNPs physically mapping between 20 Mb and 40 Mb on chromosome 6 (NCBI 36 assembly).

Trans-eQTL replication datasets

Replication of the detected eQTLs was performed in monocytes from 1,490 different samples[45] and in an independent population of 86 morbidly obese individuals that underwent elective bariatric surgery (Department of general surgery, Maastricht University Medical Centre, the Netherlands).Both these datasets also used the same Illumina HumanHT-12 expression platform.

For the 1,490 monocyte samples eQTL P-Values summary statistics were available for all monocyte *trans*-eQTLs with a nominal p < 1.0×10^{-5}. We ascertained how many of the *trans*-eQTLs we had found in our peripheral blood data had a nominal eQTL p < 1.0×10^{-5} in this monocyte dataset.

We also assessed *trans*-eQTLs in four different tissues from the 86 morbidly obese individuals that underwent bariatric surgery. DNA was extracted from blood samples using the Chemagic Magnetic Separation Module I (Chemagen) integrated with a Multiprobe II Pipeting robot (PerkinElmer). All samples were genotyped using both Illumina HumanCytoSNP-12 BeadChips and Illumina HumanOmni1-Quad BeadChips (QC was identical as was applied to the peripheral blood samples). We imputed HapMap 2 genotypes using Impute version 2.0. In addition expression profiling was performed for four different tissues for each of these individuals using the Illumina HumanHT-12 arrays. Wedge biopsies of liver, visceral adipose tissue (VAT, omentum majus), subcutaneous adipose tissue (SAT, abdominal), and muscle (musculus rectus abdominis) were taken during surgery. RNA was isolated using the Qiagen Lipid Tissue Mini Kit (Qiagen, UK, 74804). Assessment of RNA quality and concentration was done with an Agilent Bioanalyzer (Agilent Technologies USA). Starting with 200 ng of RNA, the Ambion Illumina TotalPrep Amplification Kit was used for anti-sense RNA synthesis, amplification, and purification according to the protocol provided by the manufacturer (Ambion, USA). 750 ng of complementary RNA was hybridized to Illumina HumanHT12 BeadChips and scanned on the Illumina BeadArray Reader. Expression data preprocessing was as mentioned before. We first attempted to replicate the trait-associated *trans*-eQTLs per tissue, using an FDR of 0.05 and 100 permutations. Subsequently we conducted a meta-analysis, combining the four tissues. Per *trans*-eQTL we used a weighted Z-method to combine the four individual p-values. However, these four datasets are not independent, as they reflect the same individuals. We resolved this by conducting the permutations in such a way that in every permutation round the samples were permuted in exactly the same way for each of the four tissues. By doing this we retained the correlations that exist between the different tissues per sample, and were able to get a realistic empiric (null-) distribution of expected test-statistics.

Convergence analysis

Per trait we assessed all the SNPs that have been reported to be associated with that particular trait. We analyzed per trait all possible SNP-pairs. If a pair of SNPs was not in LD ($r2 < 0.001$) we assessed whether they affected the same gene in *cis* or *trans*. When using the trait-associated *cis*- and *trans*-eQTLs that had been identified when controlling the FDR at 0.05, we identified 7 unique pairs of SNPs that caused both the same phenotype and also affected the same gene(s). When using a somewhat more relaxed set of *trans*-eQTLs, identified when controlling the FDR at 0.5, we identified 18 unique pairs of SNPs that affect the same downstream gene.

We assessed whether these numbers were significantly higher than expected, by using the same strategy that we had used to assess the enrichment of trait-associated SNPs and the HLA; we ran 100 permutations. We kept per permutation the *cis*-eQTL list as it was, but generated a permuted set of *trans*-eQTLs, equal in size to the original set of non-permuted *trans*-eQTLs. This enabled us to determine per permutation round how many unique pairs of SNPs converge on the same gene(s). We subsequently fitted a generalized extreme value distribution,

permitting us to estimate realistic enrichment significance estimates.

Co-expression between genes, based on HT12 peripheral blood co-expression

If a particular SNP is *cis*-or *trans*-acting on multiple genes, it is plausible that those genes are biologically related. Co-expression between these genes provides circumstantial evidence this is the case, strengthening the likelihood such *cis*- and *trans*-eQTLs are real. We assessed this in the peripheral blood data, by using the expression data of the 1,240 samples, run on the comprehensive HT12 expression platform. As we had removed 25 PCs (to remove physiological, environmental variation, and systematic experimental variation) for the *trans*-eQTL analyses, we decided to confine co-expression analyses to this expression dataset. As there are 43,202 HT12 probes that we mapped to a known genomic location, 43,202 * 43,201/2 = 933,184,801 probe-pairs exist. Given 1,240 samples, a Pearson correlation coefficient $r \leq 0.19$ corresponds to a $p < 0.05$ when applying stringent Bonferroni correction for these number of probe-pairs.

Accession numbers

Expression data for both the peripheral blood and the four non-blood datasets have been deposited in GEO with accession numbers GSE20142 (1,240 peripheral blood samples, hybridized to HT12 arrays), GSE20332 (229 peripheral blood samples, hybridized to H8v2 arrays) and GSE22070 (subcutaneous adipose, visceral adipose, muscle and liver samples).

Acknowledgments

We like to thank Jackie Senior for critically reading the manuscript. Furthermore, we thank all individuals for participating in this study.

Author Contributions

Conceived and designed the experiments: L.F. and G.J.t.M.

Performed the experiments: M.P., A.S., C.G.J.S., M.G.M.W, H.J.M.G., L.H.vd.B., R.A.O. and R.K.W.

Analyzed the data: R.S.N.F., L.F., G.J.t.M., H-J.W., R.C.J., J.H.V., A.S.

Contributed reagents/materials/analysis tools: R.S.N.F., H-J.W., G.J.t.M., R.C.J., L.F., C.W., D.A.v.H., L.H.vd.B., R.A.O., R.K.W., M.H.H., D.A., M.J.B., P.D., J.F., R.M.W.H., W.A.B., S.R., A.Z., C.C.E., E.M.F. and G.T.

Wrote the paper: R.S.N.F., L.F., G.J.t.M., R.C.J., C.W., M.H.H. and D.A.v.H.

Supporting information

Figure S1
Detected *cis*- and *trans*-eQTLs in genome-wide analysis.

Figure S2
Detected *cis*- and *trans*-eQTLs for 1,167 traitassociated SNPs.

Figure S3
Detected *cis*- and *trans*-eQTLs per complex trait. Immune-related and hematological associated SNPs often affect gene expression in *cis* or *trans*.

Figure S4
Co-expression distribution between eQTL genes for mean platelet volume and mean corpuscular volume.

Figure S5
Replication of *trans*-eQTLs in four non-blood tissues.

Figure S6
Principal components used as covariates in analyses.

Figure S7
Effect of removing principal components from expression data on detect ability of *cis*-eQTLs

Figure S8
Significance of detected *cis*-eQTLs before and after removal of principal components from expression data.

Figure S9
Effect of removing principal components from expression data on detect ability of *trans*-eQTLs.

Table S1
Detected *cis*-eQTLs (FDR 0.05) for all common SNPs.

Table S2
Detected *trans*-eQTLs (FDR 0.05) for all common SNPs.

Table S3
Detected *cis*-eQTLs (FDR 0.05) for 1,167 traitassociated SNPs.

Table S4
Detected *trans*-eQTLs (FDR 0.05) for 1,167 trait-associated SNPs.

Table S5
Detected *cis*- and *trans*-eQTLs (FDR 0.05) per complex trait.

Table S6
Plots of detected *trans*-eQTLs for 1,167 trait-associated SNPs for each of the seven individual cohorts of samples that make up the total of 1,469 peripheral blood samples.

Table S7
Detected *trans*-eQTLs (FDR 0.50) for 1,167 trait-associated SNPs.

Table S8
Replicated *trans*-eQTLs in monocyte eQTL dataset.

Table S9
Characteristics of subcutaneous adipose, visceral adipose, muscle and liver datasets.

Table S10
Replicated *trans*-eQTLs in subcutaneous adipose, visceral adipose, muscle and liver datasets.

Table S11
Characteristics of peripheral blood expression data.

Cell specific eQTL analysis without sorting cells

Manuscript in preparation

Harm-Jan Westra[1,47], Danny Arends[2,47], Tõnu Esko[3,4,5,6], Marjolein J. Peters[7,8], Claudia Schurmann[9,10], Katharina Schramm[11,12], Johannes Kettunen[13,14,15], Hanieh Yaghootkar[16], Benjamin P. Fairfax[17,18], Anand Kumar Andiappan[19], Yang Li[1,2], Jingyuan Fu[1], Juha Karjalainen[1], Mathieu Platteel[1], Marijn Visschedijk[1,20], Rinse Weersma[20], Silva Kasela[3,21], Lili Milani[3], Liina Tserel[22], Pärt Peterson[22], Eva Reinmaa[3], Albert Hofman[8,23], André G. Uitterlinden[7,8,23], Fernando Rivadeneira[7,8,23], Georg Homuth[9], Astrid Petersmann[24], Roberto Lorbeer[25], Holger Prokisch[26,27], Thomas Meitinger[26,27,28,29], Christian Herder[30,31], Michael Roden[30,31,32], Harald Grallert[33], Samuli Ripatti[34,35,36,37], Markus Perola[3,15], Andrew R Wood[16], David Melzer[38], Luigi Ferrucci[39], Andrew B Singleton[40], Dena G. Hernandez[40,41], Julian C. Knight[17], Rossella Melchiotti[19,42], Bernett Lee[19], Michael Poidinger[19], Francesca Zolezzi[19], Anis Larbi[19], De Yun Wang[43], Leonard H. van den Berg[44], Jan H. Veldink[44], Olaf Rotzschke[19], Seiko Makino[17], Timouthy M. Frayling[16], Veikko Salomaa[15], Konstantin Strauch[45,46], Uwe Völker[9], Joyce B.J. van Meurs[7,8], Andres Metspalu[3], Cisca Wijmenga[1], Ritsert C. Jansen[2,48], Lude Franke[1,48,49]

4

1 University of Groningen, University Medical Center Groningen, Department of Genetics, Hanzeplein 1, 9700RB, Groningen, The Netherlands
2 Groningen Bioinformatics Centre, University of Groningen, P.O. Box 11103, 9700 CC Groningen, The Netherlands
3 Estonian Genome Center, University of Tartu, Riia 23b, 51010 Tartu, Estonia.
4 Divisions of Endocrinology, Boston Children's Hospital, Boston, 02115, USA
5 Department of Genetics, Harvard Medical School, Boston, 02115, USA
6 Broad Institute, Cambridge, 02142, USA
7 Department of Internal Medicine, Erasmus Medical Centre Rotterdam, the Netherlands
8 The Netherlands Genomics Initiative-sponsored Netherlands Consortium for Healthy Aging (NGI-NCHA), Leiden / Rotterdam, the Netherlands
9 Interfaculty Institute of Genetics and Functional Genomics, University Medicine Greifswald, Friedrich-Ludwig-Jahn-Str. 15A, 17475 Greifswald, Germany
10 The Charles Bronfman Institute for Personalized Medicine, Genetics of Obesity & Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA
11 Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
12 Institut für Humangenetik, Technische Universität München, Trogerstr. 32, 81675 München, Germany
13 Computational Medicine, Institute of Health Sciences, Faculty of Medicine, University of Oulu, Oulu, Finland
14 Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland
15 Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland
16 Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK
17 Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK
18 Department of Oncology, Cancer and Haematology Centre, Churchill Hospital, Oxford, OX3 7LJ
19 Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), 8A Biomedical Grove, Singapore 138648
20 University of Groningen, University Medical Center Groningen, Department of Gastroenterology and Hepatology, Hanzeplein 1, 9700RB, Groningen, The Netherlands
21 Institute of Molecular and Cell Biology, University of Tartu, Riia 23, 51010 Tartu, Estonia
22 Molecular Pathology, Institute of Biomedicine and Translational Medicine, University of Tartu, Ravila 19, Biomedicum, 50411, Tartu, Estonia
23 Department of Epidemiology, Erasmus Medical Center Rotterdam, the Netherlands
24 Institute for Clinical Chemistry and Laboratory Medcine, University Medicine Greifswald, Sauerbruchstr., 17475 Greifswald, Germany
25 Institute for Community Medicine, University Medicine Greifswald, Walther-Rathenau-Str. 48, 17475 Greifswald, Germany
26 Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
27 Institut für Humangenetik, Technische Universität München, Trogerstr. 32, 81675 München, Germany
28 Munich Heart Allience, Munich, Germany
29 German Center for Cardiovascular Research (DZHK), Germany
30 Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Auf'm Hennekamp 65, 40225 Düsseldorf, Germany
31 German Center for Diabetes Research (DZD), partner site Düsseldorf, Germany
32 Department of Diabetology and Endocrinology, University Hospital Düsseldorf, Heinrich Heine University, Moorenstr. 5, 40225 Düsseldorf, Germany
33 Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
34 Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland
35 Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland
36 Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom
37 Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland
38 Institute of Biomedical and Clinical Sciences, University of Exeter Medical School, Barrack Road, Exeter, EX2 5DW, UK
39 Clinical Research Branch, National Institute on Aging NIAASTRA Unit, Harbor Hospital, MD, USA
40 Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, 35 Lincoln Drive, Bethesda, MD, USA
41 Department of Molecular Neuroscience and Reta Lila Laboratories, Institute of Neurology, UCL, Queen Square House, Queen Square, London WC1N 3BG, UK
42 Doctoral School in Translational and Molecular Medicine (DIMET), University of Milano-Bicocca, Piazza della Scienza, 3, 20126 Milan, Italy
43 Department of Otolaryngology, National University of Singapore, Singapore
44 Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Utrecht, The Netherlands
45 Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
46 Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
47 These authors contributed equally to this work
48 These authors jointly directed this work
49 Corresponding author (lude@ludesign.nl)

Abstract

The functional consequences of trait associated SNPs are often investigated using expression quantitative trait locus (eQTL) mapping. While trait-associated variants may operate in a cell-type specific manner, eQTL datasets for such cell-types may not always be available. We performed a genome-environment interaction (GxE) meta-analysis on data from 5,683 samples to infer the cell type specificity of whole blood *cis*-eQTLs. We demonstrate that this method is able to predict neutrophil and lymphocyte specific *cis*-eQTLs and replicate these predictions in independent cell-type specific datasets. Finally, we show that SNPs associated with Crohn's disease preferentially affect gene expression within neutrophils, including the archetypal *NOD2* locus.

Author summary

Many variants in the genome, including variants associated with disease, affect the expression of genes. These so-called expression quantitative trait loci (eQTL) can be used to gain insight in the downstream consequences of disease. While it has been shown that many disease associated variants alter gene expression in a cell-type dependent manner, eQTL datasets for specific cell types may not always be available and their sample size is often limited. We present a method that is able to detect cell type specific effects within eQTL datasets that have been generated from whole tissues (which may be composed of many cell types), in our case whole blood. By combining numerous whole blood datasets through meta-analysis, we show that we are able to detect eQTL effects that are specific for neutrophils and lymphocytes (two blood cell types). Additionally, we show that the variants associated with some diseases may preferentially alter the gene expression in one of these cell types. We conclude that our method is an alternative method to detect cell type specific eQTL effects, that may complement generating cell type specific eQTL datasets and that may be applied on other cell types and tissues as well.

Introduction

In the past seven years, genome-wide association studies (GWAS) have identified thousands of genetic variants that are associated with human disease[1]. The realization that many of the disease-predisposing variants are non-coding and that single nucleotide polymorphisms (SNPs) often affect the expression of nearby genes (i.e. *cis*-expression quantitative trait loci; *cis*-eQTLs)[2] suggests these variants have a predominantly regulatory function. Recent studies have shown that disease-predisposing variants in humans often exert their regulatory effect on gene expression in a cell-type dependent manner[3–5]. However, most human eQTL studies have used sample data obtained from mixtures of cell types (e.g. whole blood) or a few specific cell types (e.g. lympoblastoid cell lines) due to the prohibitive costs and labor required to purify subsets of cells from large samples (by cell sorting or laser capture micro-dissection). In addition, the method of cell isolation can trigger uncontrolled processes in the cell, which can cause biases.

1 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9362–7 (2009).

2 Fehrmann, R. S. N. *et al. Trans*-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA. *PLoS Genet.* 7, 14 (2011).

3 Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative Modeling of eQTLs and *Cis*-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS Genet.* 9, e1003649 (2013).

4 Fairfax, B. P. B. *et al.* Genetics of gene expression in primary immune cells identifies cell-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44, 502–510 (2012).

5 Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* 8, e1002431 (2012).

In consequence, it has been difficult to identify in which cell types most disease-associated variants exert their effect.

Here we describe a generic approach that uses eQTL data in mixtures of cell types to infer cell-type specific eQTLs (Figure 1). Our strategy includes: (i) collecting gene expression data from an entire tissue; (ii) predicting the abundance of its constituent cell types (i.e. the cell counts), by using expression levels of genes that serve as proxies for these different cell types (since not all datasets might have actual constituent cell count measurements). We used an approach similar to existing expression and methylation deconvolution methods[6–11]; (iii) run an association analysis with a term for interaction between the SNP and the proxy for cell count to detect cell-type-mediated or -specific associations, and (iv) test whether known disease associations are enriched for SNPs that show the cell-type-mediated or -specific effects on gene expression (i.e. eQTLs).

6
Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13, 86 (2012).

7
Houseman, E. A., Molitor, J. & Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* (2014).

8
Accomando, W. P., Wiencke, J. K., Houseman, E. A., Nelson, H. H. & Kelsey, K. T. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biol.* 15, R50 (2014).

9
Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* 15, R31 (2014).

10
Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–35 (2007).

11
Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat. Methods* 7, 287–9 (2010).

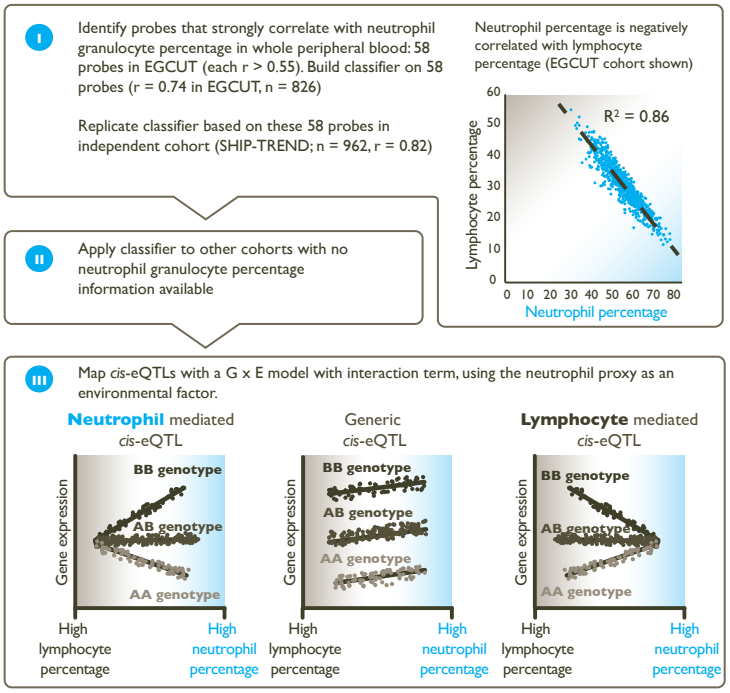


Figure 1. Method overview

I) Starting with a dataset that has cell count measurements, determine a set of probes that have a strong positive correlation to the cell count measurements. Calculate the correlation between these specific probes in the other datasets, and apply principal component analysis to combine them into a single proxy for the cell count measurement. II) Apply the prediction to other datasets lacking cell count measurements. III) Use the proxy as a covariate in a linear model with an interaction term in order to distinguish cell-type-mediated from non-cell-type-mediated eQTL effects.

Results

We applied this strategy to 5,863 unrelated, whole blood samples from seven cohorts: EGCUT[12], InCHIANTI[13], Rotterdam Study[14], Fehrmann[2], SHIP-TREND[15], KORA F4[16], and DILGOM[17]. Blood contains many different cell types that originate from either the myeloid (e.g. neutrophils and monocytes) or lymphoid lineage (e.g. B-cells and T-cells). Even though neutrophils comprise ~62% of all white blood cells, no neutrophil eQTL data have been published to date, because this cell type is particularly difficult to purify or culture in the lab[18].

For the purpose of illustrating our cell-type specific analysis strategy in the seven whole blood cohorts, we focused on neutrophils. Direct neutrophil cell counts and percentages were only available in the EGCUT and SHIP-TREND cohorts, requiring us to infer neutrophil percentages for the other five cohorts. We used the EGCUT cohort as a training dataset to identify a list of 58 Illumina HT12v3 probes that correlated positively with neutrophil percentage (Spearman's correlation coefficient $R > 0.55$). We then summarized the gene expression levels of these 58 individual probes into a single neutrophil percentage estimate, by applying principal component analysis (PCA) and using the first principal component; an approach that is similar to existing expression and methylation deconvolution methods[6-11]. We then used this procedure in the other cohorts to predict the neutrophil percentage (see Figure 2 for confirmation of the accuracy of prediction in the SHIP-TREND cohort; Spearman $R = 0.81$).

Here we limit our analysis to 13,124 *cis*-eQTLs that were previously discovered in a whole blood eQTL meta-analysis of a comparable sample size[19] (we note that these 13,124 *cis*-eQTLs were detected while assuming a generic effect across cell-types, and as such, genome-wide application of cell-type specificity strategy might result in the detection of additional cell-type-specific *cis*-eQTLs). To infer the cell-type specificity of each of these eQTLs, we performed the eQTL association analysis with a term for interaction between the SNP marker and the proxy for cell count within each cohort, followed by a meta-analysis of the

12 Metspalu, A. The Estonian Genome Project. *Drug Dev. Res.* 62, 97–101 (2004).

13 Tanaka, T. *et al.* Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet.* 5, e1000338 (2009).

14 Hofman, A. *et al.* The Rotterdam Study: 2014 objectives and design update. *Eur. J. Epidemiol.* 28, 889–926 (2013).

15 Völzke, H. *et al.* Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* 40, 294–307 (2011).

16 Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* 21, 48–54 (2013).

17 Inouye, M. *et al.* An immune response network associated with blood lipid levels. *PLoS Genet.* 6, e1001113 (2010).

18 Grisham, M. B., Engerson, T. D., McCord, J. M. & Jones, H. P. A comparative study of neutrophil purification and function. *J. Immunol. Methods* 82, 315–20 (1985).

19 Westra, H.-J. *et al.* Systematic identification of trans-eQTLs as putative drivers of known disease associations. *Nat. Genet.* (2013). doi:10.1038/ng.2756

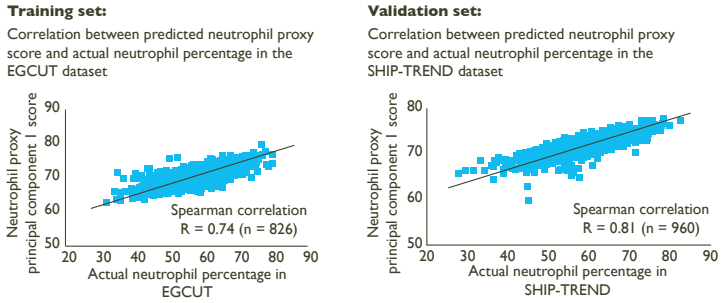


Figure 2. Validation of neutrophil proxy

There is a strong correlation between the neutrophil proxy and the actual neutrophil percentage measurements in the training dataset (EGCUT, r = 0.74). Validation of neutrophil prediction in the SHIP-TREND cohort shows a strong correlation (r = 0.81) between the neutrophil proxy and actual neutrophil percentage measurements in this dataset.

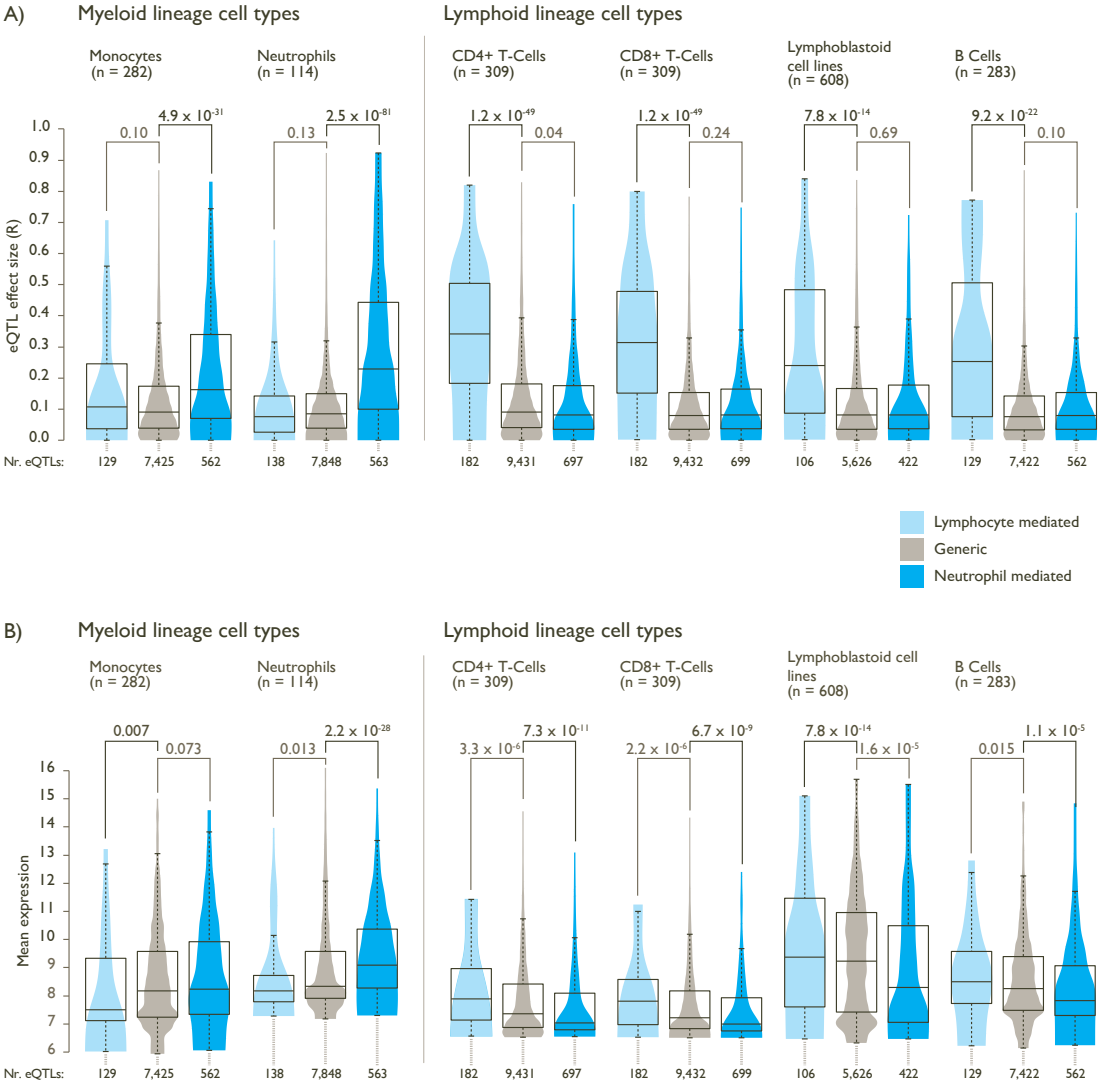


Figure 3. Validation of neutrophil and lymphoid specific *cis*-eQTLs in purified cell type eQTL datasets

A) We validated the neutrophil- and lymphoid-mediated *cis*-eQTL effects in four purified cell type datasets from the lymphoid lineage (B-cells, CD4+ T-cells, CD8+ T-cells and lymphoblastoid cell lines) and in two datasets from the myeloid lineage (monocytes and neutrophils). Compared to generic *cis*-eQTLs, large effect sizes were observed for neutrophil-mediated *cis*-eQTLs in myeloid lineage cell types, and small effect sizes in the lymphoid datasets. Conversely, lymphoid-mediated *cis*-eQTL effects had large effect sizes specifically in the lymphoid lineage datasets, while having smaller effect sizes in myeloid lineage datasets. These results indicate that our method is able to reliably predict whether a specific *cis*-eQTL is mediated by cell type.

B) Comparison between average gene expression levels between different purified cell type eQTL datasets shows that neutrophil mediated *cis*-eQTLs have, on average a lower expression in cell types derived from the lymphoid lineage, and a high expression in myeloid cell types, while the opposite is true for lymphocyte mediated *cis*-eQTLs.

interaction term (weighted for sample size) across all the cohorts. We identified 1,117 *cis*-eQTLs with a significant interaction effect (8.5% of all *cis*-eQTLs tested; false discovery rate (FDR) < 0.05; 1,037 unique SNPs and 836 unique probes; Supplementary Tables 1 and 2). Out of the total number of *cis*-eQTLs tested, 909 (6.9%) had a positive direction of effect, which indicates that these *cis*-eQTLs show stronger effect sizes when more neutrophils are present (i.e. 'neutrophil-mediated *cis*-eQTLs'; 843 unique SNPs and 692 unique probes). Another 208 (1.6%) had a negative direction of effect (196 unique SNPs and 145 unique probes), indicating a stronger *cis*-eQTL effect size when more lymphoid cells are present (i.e. 'lymphocyte-mediated *cis*-eQTLs'; since lymphocyte percentages are strongly negatively correlated with neutrophil percentages, Figure 1). Overall, the directions of the significant interaction effects were consistent across the different cohorts, indicating that our findings are robust (Supplementary Figure 1).

We validated the neutrophil- and lymphoid-mediated associations we detected in six small, purified cell-type gene expression datasets that had not been used in our meta-analysis. We generated new eQTL data from two lymphoid cell types (CD4+ and CD8+ T-cells) and one myeloid cell type (neutrophils, see online methods) and used previously generated eQTL data on two lymphoid cell types (lymphoblastoid cell lines and B-cells) and another myeloid cell type (monocytes, Supplementary Table 3). As expected, compared to *cis*-eQTLs without a significant interaction term ('generic *cis*-eQTLs', n = 12,007) the 909 neutrophil-mediated *cis*-eQTLs did indeed show very strong *cis*-eQTL effects in both of the myeloid datasets (Wilcoxon P-value ≤ 4.9×10^{-31}), and small effect sizes in the lymphoid datasets. Conversely, the 208 lymphoid-mediated *cis*-eQTLs had a pronounced effect in each of the lymphoid datasets (Wilcoxon P-value ≤ 7.8×10^{-14}; Figure 3A), while having small effect sizes in the myeloid datasets. These results indicate that our method is able to reliably predict whether a *cis*-eQTL is mediated by a specific cell type. Unfortunately, the cell type that mediates the *cis*-eQTL is not necessarily the one in which the *cis*-gene has the highest expression (Figure 3B), making it impossible to identify cell-type-specific eQTLs directly on the basis of expression levels.

Myeloid and lymphoid blood cell types provide crucial immunological functions. Therefore, we assessed five immune-related diseases for which genome-wide association studies previously identified at least 20 loci with a *cis*-eQTL in our meta-analysis. We observed a significant enrichment only for Crohn's disease (CD), (binomial test, one-tailed P = 0.002, Supplementary Table 4): out of 49 unique CD-associated SNPs showing a *cis*-eQTL effect, 11 (22%) were neutrophil-mediated. These 11 SNPs affect the expression of 14 unique genes (ordered by size of interaction effect: *IL18RAP, CPEB4, RP11-514O12.4, RNASET2, NOD2, CISD1, LGALS9, AC034220.3, SLC22A4, HOTAIRM2, ZGPAT, LIME1, SLC2A4RG*, and *PLCL1*). CD is a chronic inflammatory disease of the intestinal tract. While impaired T-cell responses and defects in antigen presenting cells have been implicated in the pathogenesis of CD, so far little attention has been paid to the role of neutrophils, because its role in the development and maintenance of intestinal inflammation is controversial: homeostatic regulation of the intestine is complex and both a depletion and an increase in

neutrophils in the intestinal submucosal space can lead to inflammation. On the one hand, neutrophils are essential in killing microbes that translocate through the mucosal layer. The mucosal layer is affected in CD, but also in monogenic diseases with neutropenia and defects in phagocyte bacterial killing, such as chronic granulomatous disease, glycogen storage disease type I, and congenital neutropenia, leading to various CD phenotypes[20]. On the other hand, an increase in activated neutrophils that secrete pro-inflammatory chemokines and cytokines (including *IL18RAP* which has a neutrophil specific eQTL) maintains inflammatory responses. Pharmacological interventions for the treatment of CD have been developed to specifically target neutrophils and *IL18RAP*, including Sagramostim[21] and Natalizumab[22]. Our new analysis shows clear neutrophil-mediated eQTL effects for many of the known CD genes, including the archetypal *NOD2* gene, and our results provide deeper insight into the role of neutrophils in CD pathogenesis.

20 Uhlig, H. H. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. *Gut* 62, 1795–805 (2013).

21 Korzenik, J. R., Dieckgraefe, B. K., Valentine, J. F., Hausman, D. F. & Gilbert, M. J. Sargramostim for active Crohn's disease. *N. Engl. J. Med.* 352, 2193–201 (2005).

22 Ghosh, S. *et al.* Natalizumab for active Crohn's disease. *N. Engl. J. Med.* 348, 24–32 (2003).

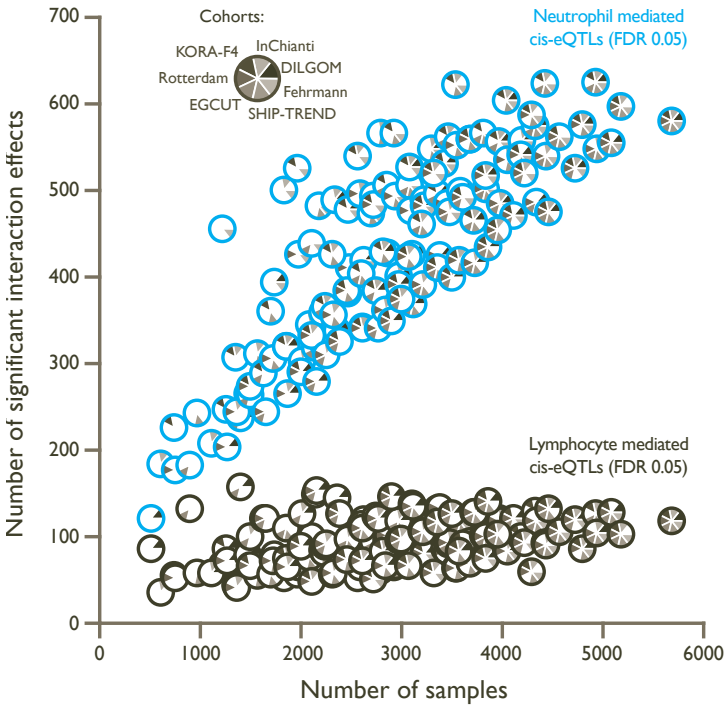


Figure 4. Effect of sample size on power to detect cell type specific *cis*-eQTLs

We systematically excluded datasets from our meta-analysis in order to determine the effect of sample size on our ability to detect significant interaction effects. The number of significant interaction effects was rapidly reduced when the sample size was decreased (the number of unique significant probes given a Bonferroni corrected P-value < 8.1×10^{-6} is shown). In general, due to their low abundance in whole blood, lymphoid-mediated *cis*-eQTL effects are harder to detect than neutrophil-mediated *cis*-eQTL effects.

Large sample sizes are essential in order to find cell-type-mediated *cis*-eQTLs (Figure 4): when we repeat our study on fewer samples (ascertained by systematically excluding more cohorts from our study), the number of significant cell-type-mediated eQTLs decreased rapidly. This was particularly important for the lymphoid-mediated *cis*-eQTLs, because myeloid cells are approximately twice as abundant as lymphoid cells in whole blood. Consequently, detecting lymphoid-mediated *cis*-eQTLs is more challenging than detecting neutrophil-specific *cis*-eQTLs. As whole blood eQTL data is easily collected, we were able to gather a sufficient sample size in order to detect cell-type-mediated or -specific associations without requiring the actual purification of cell types.

Discussion

Here we have shown that it is possible to infer in which blood cell-types *cis*-eQTLs are operating from a whole blood dataset. Cell-type proportions were predicted and subsequently used in a G x E interaction model. Hundreds of *cis*-eQTLs showed stronger effects in myeloid than lymphoid cell-types and vice versa.

These results were replicated in 6 individual purified cell-type eQTL datasets (two reflecting the myeloid and four reflecting the lymphoid lineage). This indicates our G x E analysis provides important additional biological insights for many SNPs that have previously been found to be associated with complex (molecular) traits.

Here, we concentrated on identifying *cis*-eQTLs that are preferentially operating in either myeloid or lymphoid cell-types. We did not attempt to assess this for specialized cell-types within the myeloid or lymphoid lineage. However, this is possible if cell-counts are available for these cell-types, or if these cell-counts can be predicted by using a proxy for those cell-counts. As such, identification of cell-type mediated eQTLs for previously unstudied cell-types is possible, without the need to generate new data. However, it should be noted that these individual cell-types typically have a rather low abundance within whole blood (e.g. natural killer cells only comprise ~2% of all circulating white blood cells). As a consequence, in order to have sufficient statistical power to identify eQTLs that are mediated by these cell-types, very large whole blood eQTL sample-sizes are required, specific cell types should be variable between individuals (which is analogous to the difference in the number of identified lymphoid mediated *cis*-eQTLs, as compared to the number of neutrophil mediated *cis*-eQTLs, which is likely caused by their difference in abundance in whole blood).

We confined our analyses to a subset of *cis*-eQTLs for which we had previously identified a main effect in whole peripheral blood[19]: for each *cis*-eQTL gene, we only studied the most significantly associated SNP. Considering that for many *cis*-eQTLs multiple, unlinked SNPs exist that independently affect the gene expression levels, it is possible that we have missed myeloid or lymphoid mediation of these secondary *cis*-eQTLs.

The method we have applied to predict the neutrophil percentage in the seven whole blood datasets involves correlation of gene expression probes to cell count abundances and subsequent combination of gene expression probes into a single predictor using PCA. This approach is comparable to other deconvolution methods for methylation and gene expression data[6-11]. Although we have shown that the proxy that is created by our method is able to predict neutrophil percentage accurately, this may not be the case for all cell types available in whole blood, which may be greatly dependent upon the ability of individual gene expression probes to differentiate between cell types.

However, we anticipate that the (pending) availability of large RNA-seq based eQTL datasets, statistical power to identify cell-type mediated eQTLs using our approach will improve: since RNA-seq enables very accurate gene expression level quantification and is not limited to a set of preselected probes that interrogate well known genes (as is the case for microarrays), the detection of genes that can serve as reliable proxies for individual cell-types will improve. Using RNA-seq data, it is also possible to assess whether SNPs that affect the expression of non-coding transcripts, affect splicing[23] or result in alternative polyadenylation[24] are mediated by specific cell-types.

Although we applied our method to whole blood gene expression data, our method can be applied to any tissue, alleviating the need to sort cells or to perform laser capture micro dissection. The only prerequisite for our method is the availability of a relatively small training dataset with cell count measurements in order to develop a reliable proxy for cell count measurements. Since the number of such training datasets is rapidly increasing and meta-analyses have proven successful[2,19], our approach provides a cost-effective way to identify cell-type-mediated or -specific associations that can supplement results obtained from purified cell type specific datasets, and it is likely to reveal major biological insights.

Materials & Methods

Setup of study

This eQTL meta-analysis is based on gene expression intensities measured in whole blood samples. RNA was isolated with either PAXgene Tubes (Becton Dickinson and Co., Franklin Lakes, NJ, USA) or Tempus Tubes (Life Technologies). To measure gene expression levels, Illumina Whole-Genome Expression Beadchips were used (HT12-v3 and HT12-v4 arrays, Illumina Inc., San Diego, USA). Although different identifiers are used across these different platforms, many probe sequences are identical. Meta-analysis could thus be performed if probe-sequences were equal across platforms. Integration of these probe sequences was performed as described before[19]. Genotypes were harmonized using HapMap2-based imputation using the Central European population[25]. In total, the eQTL genotype x environment interaction meta-analysis was performed on seven independent cohorts, comprising a total of 5,863 unrelated individuals. Mix-ups between gene expression samples and genotype samples were corrected using *MixupMapper*[26]. Gene expression normalization

23 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–11 (2013).

24 Zhernakova, D.V *et al.* DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* 9, e1003594 (2013).

25 The International HapMap Consortium. The International HapMap Project. 426, 789–796 (2003).

was performed as described before[19], removing up to 40 principal components (PCs). Additionally, we corrected for possible confounding factors due to arrays of poor RNA quality, by correlating the sample gene expression measurements against the first PC that was determined from the sample correlation matrix. Samples with a correlation < 0.9 were removed from further analysis.

Gene expression normalization

Each cohort performed gene expression normalization individually: gene expression data was quantile normalized to the median distribution then $\log_2$ transformed. The probe and sample means were centered to zero. Gene expression data was then corrected for possible population structure by removing four multi-dimensional scaling components (MDS components obtained from the genotype data using PLINK) using linear regression. Because normalized gene-expression data still contains large amounts of non-genetic variation[2,27], principal component analysis (PCA) was performed on the sample correlation matrix, and up to 40 principal components (PCs) were then removed from the gene expression data using linear regression[19].

In order to improve statistical power to detect cell-type mediated eQTLs, we corrected the gene expression for technical and batch effects (here we applied principal component analysis and removed per cohort the 40 strongest principal components that affect gene expression). Such procedures are commonly used when conducting *cis*-eQTL mapping[2,5,19,23,24,28]. To minimize the amount of genetic variation removed by this procedure, we performed QTL mapping for each principal component, to ascertain whether genetic variants could be detected that affected the PC. If such an effect was detected, we did not correct the gene expression data for that particular PC[19]. We chose to remove 40 PCs based on our previous study results, which suggested that this was the optimum for detecting eQTLs[19]. We would like to stress that while PC-corrected gene expression data was then used as the outcome variable in our gene x environment interaction model, we used gene expression data that was not corrected for PCs to initially create the neutrophil cell percentage proxy.

Creating a proxy for neutrophil cell percentage from gene expression data

To be able to determine whether a *cis*-eQTL is mediated by neutrophils, we reasoned that such a *cis*-eQTL would show a larger effect size in individuals with a higher percentage of neutrophils than in individuals with a lower percentage. However, this required the percentage of neutrophils in whole blood to be known, and cell-type percentage measurements were not available for all of the cohorts. We therefore created a proxy phenotype that reflected the actual neutrophil percentage that would also be applicable to datasets without neutrophil percentage measurements. In the EGCUT dataset, we first quantile normalized and $\log_2$ transformed the raw expression data. We then correlated the gene expression levels of individual probes with the neutrophil percentage, and selected 58 gene expression probes showing a high positive correlation ($r^2 > 0.3$).

26 Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* 27, 2104–11 (2011).

27 Schurmann, C. *et al.* Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PLoS One* 7, e50938 (2012).

28 Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302 (2010).

In each independent cohort, we corrected for possible confounding factors due to arrays with poor RNA quality, by correlating the sample gene expression measurements against the first PC determined from the sample correlation matrix. Only samples with a high correlation ($r \geq 0.9$) were included in further analyses. Then, for each cohort, we calculated a correlation matrix for the neutrophil proxy probes (the probes selected from the EGCUT cohort). The gene expression data used was quantile normalized, $\log_2$ transformed and corrected for MDS components. Applying PCA to the correlation matrix, we then obtained PCs that described the variation among the probes selected from the EGCUT cohort. As the first PC (PC1) contributes the largest amount of variation, we considered PC1 as a proxy-phenotype for the cell type percentages.

Determining cell-type mediation using an interaction model

Considering the overlap between the cohorts in this study and our previous study, we limited our analysis to the 13,124 *cis*-eQTLs having a significant effect (false discovery rate, FDR < 0.05) in our previous study[19]. This included 8,228 unique Illumina HT12v3 probes and 10,260 unique SNPs (7,674 SNPs that showed the strongest effect per probe, and 2,586 SNPs previously associated with complex traits and diseases, as reported in the Catalog of Published Genome-Wide Association Studies 1, on 23rd September, 2013).

We defined the model for single marker *cis*-eQTL mapping as follows:

$$Y \approx I + \beta_1 * G + e$$

where Y is the gene expression of the gene, β_1 is the slope of the linear model, G is the genotype, I is the intercept with the y-axis, and e is the general error term for any residual variation not explained by the rest of the model.

We then extended the typical linear model for single marker *cis*-eQTL mapping to include a covariate as an independent variable, and captured the interaction between the genotype and the covariate using an interaction term:

$$Y \approx I + \beta_1 * G + \beta_2 * P + \beta_3 * P{:}G + e$$

where P (cell-type proxy) is the covariate, and P:G is the interaction term between the covariate and the genotype. We used gene expression data corrected for 40 PCs as the predicted variable (Y). The interaction terms were then meta-analyzed over all cohorts using a Z-score method, weighted for the sample size[29].

Multiple testing correction

Since the gene-expression data has a correlated structure (i.e. co-expressed genes) and the genotype data also has a correlated structure (i.e. linkage disequilibrium between SNPs), a Bonferroni correction would be overly stringent. We therefore first estimated the effective number of uncorrelated tests by using permuted eQTL results from our previous *cis*-eQTL meta-analysis[19].

29 Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* 18, 1368–73 (2005).

The most significant P-value in these permutations was 8.15×10^{-5}, when averaged over all permutations. As such, the number of effective tests = $0.5 / 8.15 \times 10^{-5} \approx 6134$, which is approximately half the number of correlated *cis*-eQTL tests that we conducted (=13,124). Next, we controlled the FDR at 0.05 for the interaction analysis: for a given P-value threshold in our interaction analysis, we calculated the number of expected results (given the number of effective tests and a uniform distribution) and determined the observed number of eQTLs that were below the given P-value threshold (FDR = number of expected p-values below threshold / number of observed p-values below threshold). At an FDR of 0.05, our nominal p-value threshold was 0.009 (corresponding to an absolute interaction effect Z-score of 2.61).

Cell-type specific *cis*-eQTLs and disease

For each trait in the GWAS catalog, we pruned all SNPs with a GWAS association P-value below 5×10^{-8}, using an r^2 threshold of 0.2. We only considered traits that had more than 20 significant eQTL SNPs after pruning (irrespective of cell-type mediation). Then, we determined the proportion of pruned neutrophil-mediated *cis*-eQTLs for the trait relative to all the neutrophil-mediated *cis*-eQTLs. The difference between both proportions was then tested using a binomial test.

Acknowledgments

Acknowledgements for the respective cohorts can be found in the Supplemental Material. We would like to thank Jackie Senior for carefully editing this manuscript.

Author contributions

Development of the cell type specific eQTL mapping method: H-J.W., D.A., R.C.J. and L.F.

Computational analysis and interpretation of the results: H-J.W., D.A., T.E., M.J.P., C.S., K. Schramm, J. Kettunen, J. Karjalainen, H.Y., B.P.F., S.K., R.M., B.T., M. Poidinger and R.C.J.

Reviewing and editing of the manuscript: H-J.W., D.A., T.E., M.J.P., C.S., K. Schramm, J.K., A.K.A., Y.L., J.F, M.C., R.K.Weersma, C.W., S.K., L.M., L.T., P.P., E.R., A.H., A.G.U., F.R., G.H., H.P., T.M., C.H., M.R., H.G., S.R., A.R.W., D.M., L.Ferruci, A.B.S., D.G.H, R.M., B.T., M. Poidinger, F.Z., A.L., D.Y.W., O.R., K.S., U.V., J.B.J.M., A.M., R.C.J. and L.F.

Data collection: T.E., H.Y., B.P.F., A.K.A., M. Platteel, L.M., L.T., P.P., E.R., A.H., A.G.U., A.P., R.L., H.P., T.M., C.H., M.R., H.G., M. Perola, S.M., J.C.K., D.Y.W., L.H.v.d.B, J.H.V., O.R., T.M.F., V.S., K. Strauch and A.M.

Competing interests

No competing interests have been declared by any of the participating cohorts.

Data Access

The source code and documentation for this type of analysis are available as part of the eQTL meta-analysis pipeline at https://github.com/molgenis/systemsgenetics
Summary results are available from http://www.genenetwork.nl/celltype

Accession numbers

Discovery cohorts: Fehrmann (GSE 20142), SHIP-TREND (GSE 36382), Rotterdam Study (GSE 33828), EGCUT (GSE 48348), DILGOM (E-TABM-1036), InCHIANTI (GSE 48152), KORA F4 (E-MTAB-1708).

Replication Cohorts: Stranger (E-MTAB-264), Oxford (E-MTAB-945).

Supplementary Information

Supplementary Figure 1 Comparison of effect sizes and effect direction between datasets
Comparison of interaction effect Z-scores shows a high consistent direction of effect between datasets and with the meta-analysis for those interaction effects significant at FDR < 0.05.

Supplementary Table 1
Summary statistics for the interaction analysis.

Supplementary Table 2
Results of the interaction analysis.

Supplementary Table 3
Summary statistics showing the effect size (correlation coefficient) in each of the tested replication datasets.

Supplementary Table 4
Results of the neutrophil mediated *cis*-eQTL disease enrichment analysis.

chapter 4

Cell specific eQTL analysis without sorting cells

Part

2

Human disease-associated genetic variation impacts large intergenic non-coding RNA expression

PLoS Genetics, 2013 January; 9, e1003201

Vinod Kumar[1], Harm-Jan Westra[1.], Juha Karjalainen[1.], Daria V. Zhernakova[1.], Tõnu Esko[2], Barbara Hrdlickova1, Rodrigo Almeida[1,3], Alexandra Zhernakova[1], Eva Reinmaa[2], Urmo Võsa[2], Marten H. Hofker[4], Rudolf S. N. Fehrmann[1], Jingyuan Fu[1], Sebo Withoff[1], Andres Metspalu[2], Lude Franke[1"], Cisca Wijmenga[1"*]

1 Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
2 Institute of Molecular and Cell Biology and Estonian Genome Center, University of Tartu, Tartu, Estonia
3 Graduate Program in Health Sciences, University of Brasilia School of Health Sciences, Brasilia, Brazil
4 Molecular Genetics Section, Department of Pathology and Medical Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Abstract

Recently it has become clear that only a small percentage (7%) of disease-associated single nucleotide polymorphisms (SNPs) are located in protein-coding regions, while the remaining 93% are located in gene regulatory regions or in intergenic regions. Thus, the understanding of how genetic variations control the expression of non-coding RNAs (in a tissue-dependent manner) has far-reaching implications. We tested the association of SNPs with expression levels (eQTLs) of large intergenic non-coding RNAs (lincRNAs), using genome-wide gene expression and genotype data from five different tissues. We identified 112 *cis*-regulated lincRNAs, of which 45% could be replicated in an independent dataset. We observed that 75% of the SNPs affecting lincRNA expression (lincRNA *cis*-eQTLs) were specific to lincRNA alone and did not affect the expression of neighboring protein-coding genes. We show that this specific genotype-lincRNA expression correlation is tissue-dependent and that many of these lincRNA *cis*-eQTL SNPs are also associated with complex traits and diseases.

Author Summary

Large intergenic non-coding RNAs (lincRNAs) are the largest class of non-coding RNA molecules in the human genome. Many genome-wide association studies (GWAS) have mapped disease-associated genetic variants (SNPs) to, or in, the vicinity of such lincRNA regions. However, it is not clear how these SNPs can affect the disease. We tested whether SNPs were also associated with the lincRNA expression levels in five different human primary tissues. We observed that there is a strong genotype-lincRNA expression correlation that is tissue-dependent. Many of the observed lincRNA *cis*-eQTLs are disease-or trait-associated SNPs. Our results suggest that lincRNA-eQTLs represent a novel link between non-coding SNPs and the expression of protein-coding genes, which can be exploited to understand the process of gene-regulation through lincRNAs in more detail.

Introduction

It is now evident that most of the human genome is transcribed to produce not only protein-coding transcripts but also large numbers of non-coding RNAs (ncRNAs) of different size[1,2]. Well-characterized short ncRNAs include microRNAs, small interfering RNAs, and piwi-interacting RNAs, whereas the large intergenic non-coding RNAs (lincRNAs) make up most of the long ncRNAs. LincRNAs are non-coding transcripts of more than 200 nucleotides long; they have an exon-intron-exon structure, similar to protein-coding genes, but do not encompass open-reading frames[3]. The recent description of more than 8,000 lincRNAs makes these the largest subclass of the non-coding transcriptome in humans[4].

Evidence is mounting that lincRNAs participate in a wide-range of biological processes such as regulation of epigenetic signatures and gene expression[5–7], maintenance of pluripotency and

1 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816 (2007).

2 Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–8 (2007).

3 Ørom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143, 46–58 (2010).

4 Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–27 (2011).

Table 1.

Some of the lincRNA *cis*-eQTLs are disease-associated SNPs. Chr chromosome, SAT Saturated adipose tissue, VAT Visceral adipose tissue

***Cis*-eQTL SNP**	**eQTL P on lincRNA**	**Proxies associated with disease (R2>0.8)**	**Chr**	**Trait/Disease**	**eQTL affected lincRNA**	**eQTL tissue**
rs13278062	4.31×10^{-32}	rs13278062	8	Exudative age-related macular degeneration	XLOC_006742	Blood
rs11066054	4.09×10^{-11}	rs6490294	12	Mean platelet volume	XLOC_010202	Blood
rs206942	3.63×10^{-5}	rs206936	6	Body mass index	XLOC_005690	Blood
rs11065766	6.67×10^{-5}	rs10849915	12	Alcohol consumption	XLOC_009878	Blood
	6.67×10^{-5}	rs10774610	2	Drinking behavior		
rs1465541	1.84×10^{-4}	rs11684202	2	Coronary heart disease	XLOC_002026	Blood
rs12125055	1.84×10^{-4}	rs7542900	1	Type 2 diabetes	XLOC_000922	Blood
rs199439	8.25×10^{-6}	rs199515	17	Parkinson's disease	XLOC_012496	SAT
		rs415430	17	Parkinson's disease		SAT
		rs199533	17	Parkinson's disease		SAT
rs17767419	1.05×10^{-8}	rs17767419	16	Thyroid volume	XLOC_011797	SAT, VAT
		rs3813582	16	Thyroid function		SAT, VAT

differentiation of embryonic stem cells[8]. In addition, several individual lincRNAs have also been implicated in human diseases. A well-known example is a region on chromosome 9p21 that encompasses an antisense lincRNA, *ANRIL* (antisense lincRNA of the *INK4* locus). Genome-wide association studies (GWAS) have shown that this region is significantly associated with susceptibility to type 2 diabetes, coronary disease, and intracranial aneurysm as well as different types of cancers[9] and some of the associated SNPs have been shown to alter the transcription and processing of *ANRIL* transcripts[10]. Similarly, increased expression of lincRNA *HOTAIR* (*HOX* antisense non-coding RNA) in breast cancer is associated with poor prognosis and tumor metastasis[10]. Another example is *MALAT-1* (metastasis associated in lung adenocarcinoma transcript) where the expression is three-fold higher in metastasizing tumors of non-small-cell lung cancer than in non-metastasizing tumors[11].

In addition, over the last decade, more than 1,200 GWAS have identified nearly 6,500 disease- or trait-predisposing SNPs, but only 7% of these are located in protein-coding regions[12,13]. The remaining 93% are located within non-coding regions[14], suggesting that GWAS-associated SNPs regulate gene transcription levels rather than altering the protein-coding sequence or protein structure. Even though there is growing evidence to implicate lincRNAs in human diseases[15,16], it is unknown whether disease-associated SNPs could affect the expression of non-coding RNAs. We hypothesized that GWAS-associated SNPs can affect the expression of lincRNA genes, thereby proposing a novel disease mechanism.

To test this hypothesis, we performed eQTL mapping on 2,140 human lincRNA-probes using genome-wide gene expression and genotype data of 1,240 peripheral blood samples (discovery cohort)[17]. The lincRNA *cis*-eQTLs identified were then tested for replication in an independent cohort containing 891 peripheral blood samples (replication cohort). Since lincRNAs are considered

5
Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 106, 11667–72 (2009).

6
Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–23 (2007).

7
Nagano, T. *et al.* The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 322, 1717–20 (2008).

8
Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477, 295–300 (2011).

9
Pasmant, E., Sabbagh, A., Vidaud, M. & Bièche, I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25, 444–8 (2011).

10
Burd, C. E. *et al.* Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet.* 6, e1001233 (2010).

11
Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–41 (2003).

12
Pennisi, E. The Biology of Genomes. Disease risk links to gene regulation. *Science* 332, 1031 (2011).

13
Kumar, V. *et al.* Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. *PLoS Genet.* 9, e1003201 (2013).

14
Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9362–7 (2009).

to be more tissue-specific than protein-coding genes[4], we set-out to identify tissue-dependent *cis*-eQTLs for lincRNAs using data from another four different primary tissues from the subset of 85 individuals in our primary cohort[18]. Subsequently, we tested whether SNPs that affect the levels of lincRNA expression are associated with diseases or traits. Finally, we predicted the most likely function(s) of a subset of *cis*-eQTL lincRNAs by using co-regulation information from a compendium of approximately 80,000 expression arrays (www.GeneNetwork.nl).

Results

Commercial microarrays contain probes for a subset of non-coding RNA

Whole-genome gene expression oligonucleotide arrays have played a crucial role in our understanding of gene regulatory networks. Even though most of the currently available commercial microarrays are designed to capture all known protein-coding transcripts, they still include subsets of probes that capture transcripts of unknown function (sometimes abbreviated as TUF). We investigated whether the TUF probes present on the Illumina Human HT12v3 array, overlap with lincRNA transcripts that were recently described in the lincRNA catalog[4]. The lincRNA catalog contained a provisional set of 14,393 transcripts mapping to 8,273 lincRNA genes and a stringent set of 9,918 transcripts mapping to 4,283 lincRNA genes. We identified 2,140 unique probes that map to 1,771 different lincRNAs from the provisional set and 1,325 unique probes that map to 1,051 lincRNA genes from the stringent set. We chose 2,140 unique probes that mapped to lincRNAs from the provisional set for further eQTL analysis.

Genetic control of lincRNAs expression in blood

It is known that in general lincRNAs are less abundantly expressed compared to protein-coding transcripts[4]. To test the expression levels of the 2,140 lincRNA probes in 1,240 peripheral blood samples (discovery cohort), we compared the quantile-normalized, log scale transformed mean expression intensity as well as expression variation of the lincRNA probes to probes mapping to protein-coding transcripts. We indeed observed a significant difference in the expression levels, where lincRNA probes are less abundant (mean expression = 6.67) than probes mapping to protein-coding transcripts (mean expression = 6.92, Wilcoxon Mann Whitney $P < 2.2 \times 10^{-16}$; Figure S1). We also observed a highly significant difference in the expression variation between lincRNA probes and probes mapping to protein-coding transcripts (Wilcoxon Mann Whitney $P < 3.85 \times 10^{-96}$). Next, we tested whether the expression of these 2,140 lincRNA probes is affected by SNPs in *cis*, by performing eQTL mapping in these 1,240 peripheral blood samples for which genotype data was also available. We confined our analysis to SNP-probe combinations for which the distance from the center of the probe to the genomic location of the SNP was ≤ 250 kb. In the end, at a false-discovery rate (FDR) of 0.05, we identified 5,201 significant SNP-probe combinations, reflecting 4,644 different SNPs; these affected the expression of 112 out of 2,140 different lincRNA probes. The 112 lincRNA probes mapped to 108 lincRNA genes and comprised 5.2% of all tested lincRNA probes, with a nominal significance

15 Martin, L. & Chang, H.Y. Uncovering the role of genomic "dark matter" in human disease. *J. Clin. Invest.* 122, 1589–95 (2012).

16 Jendrzejewski, J. *et al.* The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc. Natl. Acad. Sci. U. S. A.* 109, 8646–51 (2012).

17 Fehrmann, R. S. N. *et al. Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* 7, e1002197 (2011).

18 Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* 8, e1002431 (2012).

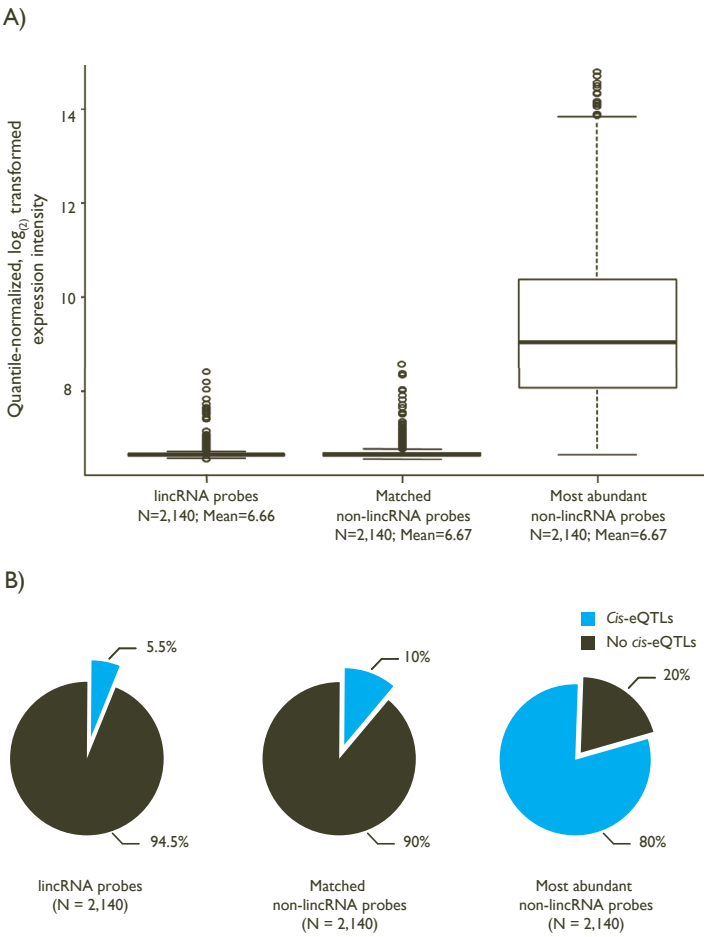


Figure 1. The number of detected *cis*-eQTLs is dependent on the expression levels of the transcripts.

A) Quantile-normalized average expression intensity and B) number of *cis*-eQTL affected probes in percentage, for 2,140 lincRNA probes, 2,140 non-lincRNA (matched for 2,140 lincRNA probes' median expression and standard deviation) and 2,140 most abundantly expressed non-lincRNA probes.

ranging from $P < 2.8 \times 10^{-4}$ to 9.81×10^{-198} in peripheral blood (Table S1).

Replication of lincRNA *cis*-eQTLs in an independent blood dataset

We then performed a replication analysis to test the reproducibility of the identified 112 lincRNA *cis*-eQTLs using an independent dataset of 891 whole peripheral blood samples. We took the 112 lincRNA-probes (or 5,201 SNP-probe pairs) that were significantly affected by *cis*-eQTLs in the discovery cohort and tested whether these eQTLs were also significant in the replication dataset (at FDR 0.05). We could replicate 45% of the 112 lincRNA *cis*-eQTLs at an FDR < 0.05, of which all the eQTLs had an identical allelic direction (Figure S2). The smaller sample size of the replication cohort compared to the discovery cohort makes it inherently difficult to replicate all the *cis*-eQTLs that we have detected in the discovery cohort.

Number of *cis*-eQTLs is dependent on expression levels of transcripts

Our observation that 5.2% of all tested lincRNAs are *cis*-regulated (Table S1) might seem disappointing, compared to our earlier observation that 25% of the protein-coding probes in this dataset are *cis*-regulated[18]. However, we reasoned that the generally lower expression levels of lincRNAs compared to protein-coding genes might make it more difficult to detect *cis*-eQTLs for lincRNAs, as the influence of background noise becomes substantial for less abundant transcripts, making accurate expression quantification difficult (Figure S1A).

Indeed, we found significantly higher expression levels for the 112 *cis*-eQTL lincRNA probes (mean expression = 6.80) compared to the 2,028 non-eQTL lincRNA probes (mean expression = 6.66 Wilcoxon Mann Whitney $P = 3.88 \times 10^{-15}$; Figure S3) and also observed a significant difference in expression variance between the 112 *cis*-eQTL lincRNAs compared to the 2,028 non-*cis*-eQTL lincRNAs (Wilcoxon Mann Whitney $P = 1.067 \times 10^{-8}$), indicating that lower overall expression levels do make identification of *cis*-eQTLs more difficult.

To further confirm the relationship between average expression levels of probes and the number of detectable *cis*-eQTLs, we first mapped *cis*-eQTLs for an equal set of 2,140 probes that were instead protein-coding and were the most abundantly expressed of all protein-coding probes. We also conducted *cis*-eQTL mapping for a set of 2,140 protein-coding probes that had been selected to have an identical expression intensity distribution as the 2,140 lincRNA probes (i.e. matched for mean expression intensity and standard deviation), using the same 1,240 blood samples (Figure 1A). We indeed observed a profound relationship between average expression levels of protein-coding transcripts and the number of detectable *cis*-eQTLs. Eighty percent of the 2,140 most abundantly expressed protein-coding probes showed a *cis*-eQTL effect, whereas only 10% of the protein-coding probes that had been matched for an expression intensity of the 2,140 lincRNA-probes were affected by *cis*-eQTLs (Figure 1B).

Hence it is possible that if we can accurately quantify all lincRNAs in large RNA-sequencing datasets, we will be able to identify *cis*-eQTLs for a larger proportion of all lincRNAs. Most SNPs that affect lincRNA expression do not alter the expression of protein-coding genes. It could be possible that the SNPs that affect lincRNA expression actually operate by first affecting protein-coding gene expression levels, which in turn affect lincRNA expression. If this were to be the case, our identified lincRNA *cis*-eQTLs would merely be a by-product of protein-coding *cis*-eQTLs. To ascertain this, we tested whether the 112 lincRNA-eQTL SNPs were also significantly affecting neighboring protein-coding genes. By keeping the same significance threshold (at FDR < 0.05 level, the value threshold was 2.4×10^{-4}), we observed that nearly 75% (83 out of 112) of the lincRNA-eQTLs were affecting only lincRNAs, even though the interrogated neighboring protein-coding genes were generally more abundantly expressed than the lincRNAs themselves (Figure S4). Genetic variants can thus directly regulate the expression levels of lincRNAs.

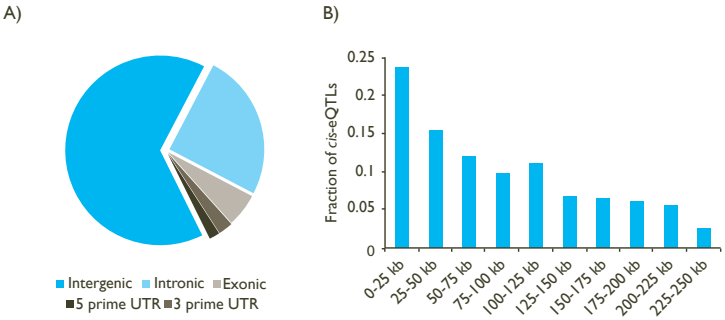


Figure 2. Distribution of lincRNA *cis*-eQTLs with respect to different transcripts.

(A) The majority of the lincRNA *cis*-eQTLs are located within the non-coding part of the genome and less than 6% of lincRNA *cis*-eQTLs are located within mRNA.

B) Distribution of lincRNA *cis*-eQTLs with respect to distance to the lincRNA transcripts. The x-axis displays the 250 kb window used for *cis*-eQTL mapping and the y-axis displays the fraction of lincRNA *cis*-eQTLs located within this window.

We found 29 *cis*-eQTLs to be associated with the expression of both lincRNA and protein coding genes. For 50% of these 29 *cis*-eQTLs, we found that the expression of lincRNAs and protein-coding genes was in the opposite direction, whereas for the other 50% of *cis*-eQTLs, both types of transcripts were co-regulated in the same direction (Figure S5). We tested whether these 29 *cis*-eQTLs are the strongest eQTLs for both lincRNA and protein-coding genes. Although these 29 *cis*-eQTLs were the strongest eQTLs for lincRNAs, only 5 among 29 were also the strongest eQTLs for protein-coding genes. This observation further highlights the direct regulation of lincRNA expression through genetic variants.

Some lincRNA *cis*-eQTLs are tissue-dependent

There is considerable interest in mapping eQTLs in disease-relevant tissue types. We reasoned that since expression of the lincRNAs seems to be much more tissue-specific than the expression of protein-coding genes[4], mapping lincRNA-eQTLs in different tissues could reveal additional, tissue-specific lincRNA-eQTLs. To test this, we analyzed gene expression and genotype data of 74 liver samples, 62 muscle samples, 83 subcutaneous adipose tissue (SAT) samples, and 77 visceral adipose tissue (VAT) samples from our primary cohort of 85 unrelated, obese Dutch individuals[18]. Upon *cis*-eQTL mapping we detected 35 *cis*-eQTL-probes, of which 18 were specific in the four different non-blood tissues, resulting in a total of 130 lincRNA-eQTLs in the combined set of all five tissues (Table S1). Five *cis*-eQTLs identified in blood tissue were also significantly replicated in at least one other non-blood tissue (Table S1). While we could replicate 45% of the *cis*-eQTLs in the substantial whole peripheral blood replication cohort, the replication rate in the very small cohorts for fat, liver and muscle tissue was, as expected, much lower. We were able to observe tissue-specific lincRNA eQTLs in muscle (1), liver (4), SAT (9) and blood (107) (Figure S6). Since the four non-blood tissue expression levels were from the same individuals, these results do indeed provide evidence that some of the lincRNAs are regulated by genetic variants in a tissue-specific manner.

LincRNA tissue specific *cis*-eQTLs are disease-associated SNPs

As most of the GWAS-associated SNPs are located within non-coding regions, we tested whether the 130 lincRNA-eQTLs identified in five different tissues are also GWAS-associated variants. To do this, we intersected trait-associated SNPs (at reported nominal $P < 9.9 \times 10^{-6}$, retrieved from the catalog of published genome-wide association studies per 26 July 2012)[14] with the 130 top lincRNA *cis*-eQTLs and their proxies (proxies with $R^2 > 0.8$ using the 1000 Genome CEU population as reference). We identified 12 GWAS SNPs or their proxies, that were also a lincRNA *cis*-eQTLs of eight different lincRNA genes (Table 1). All except one of the 12 SNPs were exclusively associated with lincRNA expression and thus did not affect the expression levels of neighboring protein-coding genes (Table 1), suggesting a causative role of altered lincRNA expression for these phenotypes. Notably SNP rs13278062 at 8p21.1, associated with exudative age-related macular degeneration (AMD) in the Japanese population, was reported to alter the transcriptional levels of *TNFRSF10A* (Tumor necrosis factor receptor superfamily 10A) protein-coding gene[19]. Here we identified SNP rs13278062 as a highly significant *cis*-eQTL of lincRNA *XLOC_006742* (*LOC389641*; $P = 4.31 \times 10^{-32}$) rather than for *TNFRSF10A* ($P = 4.21 \times 10^{-4}$) protein-coding gene (Figure S7). Furthermore, SNP rs13278062 is located in exon 1 of lincRNA *XLOC_006742*, which encompasses an ENCODE (Encyclopedia of DNA elements) enhancer region characterized by H3K27 acetylation and DNaseI hypersensitive clusters[20] (Figure S8).

Another interesting example is at 17q21.31 where three Parkinson's disease associated SNPs were in strong linkage disequilibrium ($R^2 > 0.8$) with top *cis*-eQTL SNP rs199439, which affects lincRNA *XLOC_012496* expression exclusively in SAT (Table 1). Weight loss due to body-fat wasting is a very common but poorly understood phenomenon in Parkinson's disease patients[21]. In this regard, it is intriguing to note that the Parkinson's disease associated SNPs affects lincRNA expression exclusively in fat tissue (Table 1). Hence, identifying lincRNA-eQTLs in disease-relevant tissue types using larger groups of individuals may open up new avenues towards achieving a better understanding of disease mechanisms.

LincRNA function predictions using a co-expression network of ~ 80,000 arrays: a mechanistic link between disease and lincRNA

Our observations suggest a role for lincRNAs in complex diseases and other phenotypes. The next, rather daunting task is to elucidate the function of these ncRNAs. We recently developed a co-regulation network (GeneNetwork, www.genenetwork.nl/genenetwork, manuscript in preparation), to predict the function of any transcript based on co-expression data extracted from approximately 80,000 Affymetrix microarray experiments (see Methods). We interrogated the GeneNetwork database to predict the function of eQTL-affected lincRNAs. Among the 130 *cis*-eQTL lincRNAs that we had identified in the five different tissues, 43 were represented by expression probe sets on Affymetrix arrays for which we could predict the function (Table S2). These 43 probes include four out of eight disease-associated lincRNAs described above (Table 1) and function prediction for these probes provided relevant biological explanations.

19
Arakawa, S. *et al.* Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population. *Nat. Genet.* 43, 1001–4 (2011).
20
Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* 12, 996–1006 (2002).
21
Kashihara, K. Weight loss in Parkinson's disease. *J. Neurol.* 253 Suppl , VII38–41 (2006).

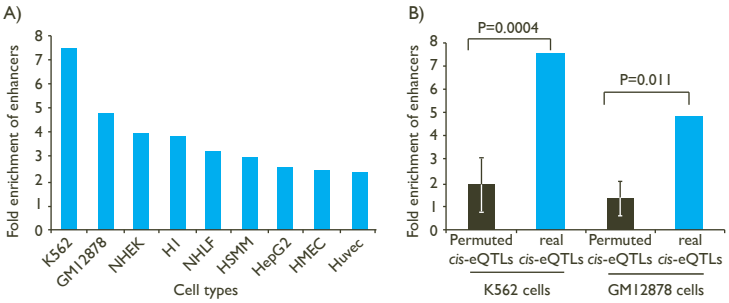


Figure 3. Localization of lincRNA *cis*-eQTLs in regulatory regions.

(A) A plot to indicate the location of lincRNA *cis*-eQTLs in cell-specific enhancers. The x-axis shows the different cell lines analyzed and the y-axis shows the fold enrichment of enhancers. (B) A plot to show the difference in fold enrichment of enhancers for real lincRNA *cis*-eQTLs compared to permuted lincRNA *cis*-eQTLs. The significance of the difference in fold enrichment was tested by T-test. The HaploReg database was used to analyze the fold enrichment of enhancers.

LincRNA co-expression analysis: disease-associated lincRNAs are co-expressed with neighboring protein-coding genes

It has been reported that some transcribed long ncRNAs function as enhancers that regulate the expression of neighboring genes[3] and may thereby contribute to the disease pathology. We found that the AMD-associated lincRNA *XLOC_006742* (*LOC389641*) (by virtue of SNP rs13278062 which exhibits a significant eQTL effect) (Figure S7) is in strong co-expression with *TNFRSF10A* based on our GeneNetwork database (Table S3). AMD is a leading cause of blindness among elderly individuals worldwide and recent studies, both in animal models and in humans, provide compelling evidence for the role of immune system cells in its pathogenesis[22]. The gene *TNFRSF10A*, which encodes TRAIL receptor 1 (TRAIL1), has been implicated as a causative gene for AMD[19]. It has been shown that binding of TRAIL to TRAILR1 can induce apoptosis through caspase 8 activation[23] and using GeneNetwork we also predict a role in apoptosis for lincRNA *XLOC_006742* (Table S2). Another trait-associated SNP, rs11065766, is the top *cis*-eQTL of lincRNA *XLOC_009878* (*ENSG00000185847* or *RP1-46F2.2* or *LOC100131138*) and it is in strong linkage disequilibrium with two SNPs associated with alcohol drinking behavior (Table 1). We found that the lincRNA *XLOC_009878* is strongly co-expressed with the neighboring protein-coding gene *MYL2* (Table S4) and, according to our predictions, lincRNA *XLOC_009878* is involved in striated muscle contraction ($P = 1.22 \times 10^{-26}$). Chronic alcohol abuse can lead to striking changes in skeletal muscle structure, which in turn plays a role in the development of alcoholic myopathy and/or cardiomyopathy[24]. It has also been reported that alcohol can reduce the content of skeletal muscle proteins such as titin and nebulin to affect muscle function in rats[25]. We found lincRNA *XLOC_009878* to be co-expressed with titin and many other skeletal muscle proteins necessary for the structural integrity of the muscle (Table S4). Thus, it needs to be tested whether deregulation of lincRNA *XLOC_009878* expression might alter an individual's ability to metabolize alcohol due to changes in the muscle functional property.

22 Patel, M. & Chan, C.-C. Immunopathological aspects of age-related macular degeneration. *Semin. Immunopathol.* 30, 97–110 (2008).

23 Johnstone, R. W., Frew, A. J. & Smyth, M. J. The TRAIL apoptotic pathway in cancer onset, progression and therapy. *Nat. Rev. Cancer* 8, 782–98 (2008).

24 George, A. & Figueredo, V. M. Alcoholic cardiomyopathy: a review. *J. Card. Fail.* 17, 844–9 (2011).

25 Hunter, R. J. *et al.* Alcohol affects the skeletal muscle proteins, titin and nebulin in male and female rats. *J. Nutr.* 133, 1154–7 (2003).

Localization of lincRNA *cis*-eQTLs in regulatory regions

We found that more than 70% of the lincRNA *cis*-eQTLs from both blood and non-blood tissues were located in intergenic regions with respect to protein-coding genes (Figure 2A). We also found high frequencies of lincRNA *cis*-eQTLs to be located around transcriptional start site (Figure 2B), suggesting that these *cis*-eQTLs may affect the expression of lincRNAs through similar gene regulatory mechanisms as those seen for protein-coding *cis*-eQTLs. Thus, in order to understand the mechanism of how lincRNA *cis*-eQTLs affect lincRNA expression, we intersected the location of top 112 lincRNA *cis*-eQTLs and their proxies ($r^2 = 1$) in blood with regulatory regions using the HaploReg database[26]. The results suggested that indeed most of the lincRNA *cis*-eQTLs (69%) were located in functionally important regulatory regions (Figure S8), which contained DNAse I regions, transcription factor binding regions, and histone marks of promoter and enhancer regions. Furthermore, these *cis*-eQTLs were found to be located more often within blood cell-specific enhancers (K562 and GM12878) (Figure 3A), suggesting that some of these *cis*-eQTLs regulate lincRNA expression in a tissue-specific manner through altering these enhancer sequences. Since we observed enrichment of cell-specific enhancers for lincRNA *cis*-eQTLs within blood cells (K562 and GM12878), we compared the fold enrichment of enhancers in these two cell types to see whether lincRNA *cis*-eQTLs are more often located in functionally important regions than any random set of SNPs. We found a significant difference in the enrichment of enhancers in which more than a 4-fold enrichment was seen for real *cis*-eQTLs both in K562 cells ($P = 0.0004$) and GM12878 cells ($P = 0.011$) compared to permuted SNPs. These findings suggest that some of the identified lincRNA *cis*-eQTLs are indeed functional SNPs.

Discussion

Even though it may have been expected that lincRNA expression would be under genetic control, this is the first study, to our knowledge, to comprehensively establish this link. We were able to identify *cis*-eQTLs in five different tissues and have demonstrated that common genetic variants regulate the expression of lincRNAs alone. It is intriguing that around 75% of lincRNA *cis*-eQTLs are specific to lincRNAs alone, but not to protein-coding genes. Recent data from the ENCODE project suggests that combinations of different transcription factors are involved in regulating gene-expression in different cell types and non-coding RNAs tend to be regulated by certain combinations of transcription factors more often than others[27]. Thus, it could still be possible that some transcription factors specifically regulate lincRNA expression. We also observed a strong relationship between whether or not a transcript is affected by *cis*-eQTLs and its expression levels, where highly abundant transcripts were more often affected by *cis*-eQTLs. This relationship was comparable between lincRNA and protein-coding probes, although protein-coding probes (matched for expression levels of lincRNA probes) tend to show more *cis*-eQTLs (Figure 1B; 5.2% versus 10%). Although this difference is not drastic, it may suggest that lincRNAs exhibit another layer of gene regulation which is more

26 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–4 (2012).

27 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100 (2012).

tissue-specific. Thus, we may expect to identify many more lincRNA *cis*-eQTLs once larger datasets of different tissues become available.

One limitation of our study is the lack of probes to comprehensively map eQTLs to all the reported lincRNAs, as we relied upon microarrays. Future analyses using RNA-sequencing datasets will undoubtedly provide much more insight into how genetic variants affect lincRNA expression. So far, two landmark RNA-sequencing based eQTL studies have been published using 60 (Montgomery *et al.*)[28] and 69 samples (Pickrell *et al.*)[29], respectively. While Pickrell *et al.* did not mention lincRNAs with a *cis*-eQTL effect, Montgomery et al identified six *cis*-regulated lincRNAs (at a slightly higher FDR of 0.17). We re-analyzed these two datasets and found that we could replicate one of the 112 *cis*-eQTL lincRNAs effects that we detected using arrays (with an identical allelic direction; Figure S10). These results indicate that *cis*-eQTL lincRNAs detected using conventional microarrays can be replicated in sequencing-based datasets. However, it also indicates that sample size is currently a limiting factor in finding many more *cis*-eQTL lincRNAs in sequencing-based datasets.

Nevertheless, our results clearly indicate that there is a strong genotype-lincRNA expression correlation that is tissue-dependent. A considerable number of the observed lincRNA *cis*-eQTLs are disease-or trait-associated SNPs. Since lincRNAs can regulate the expression of protein-coding genes either in *cis*[3] or in *trans*[8], lincRNA-eQTLs represent a novel link between non-coding SNPs and the expression of protein-coding genes. Our examples show that this link can be exploited to understand the process of gene-regulation in more detail, which may assist us in characterizing lincRNAs as another class of disease biomarkers.

Methods

Ethics statement

This study was approved by the Medical Ethical Board of Maastricht University Medical Center (four non-blood tissues), and local ethical review boards (1,240 peripheral blood samples) in line with the guidelines of the 1975 Declaration of Helsinki. Informed consent in writing was obtained from each subject personally. The subject information is provided in Table S5.

Mapping probes to lincRNAs

A detailed mapping strategy of Illumina expression probe sequences has been described previously[17]. We extracted 43,202 expression probes mapping to single genomic locations (hg18 build) and excluded those that did not map or that mapped to multiple different loci. LincRNA chromosomal coordinates (hg19 build) were obtained from the lincRNA catalog (http://www.broadinstitute.org/genome_bio/human_lincrnas/?q=lincRNA_catalog) and converted to hg18 coordinates using UCSC's LiftOver application (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Subsequently, we extracted probes mapping to lincRNA exonic regions by employing BEDtools[30].

28 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–7 (2010).
29 Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–72 (2010).

Blood dataset of 1,240 samples

The blood dataset and a detailed eQTL mapping strategy have been described previously[17]. Briefly, 1,240 peripheral blood samples from unrelated, Dutch control subjects were investigated (Table S5). Genotyping of these samples was performed according to Illumina's standard protocols (Illumina, San Diego, USA), using either the HumanHap370 or 610-Quad platforms. Because the non-blood samples (see below) were genotyped using Illumina HumanOmni1-Quad BeadChips, we applied IMPUTE v2[31] to impute the genotypes of SNPs that were covered by the Omni1-Quad chip but that were not included on the Hap370 or 610-Quad platforms[31]. Anti-sense RNA was synthesized using the Ambion Illumina TotalPrep Amplification Kit (Ambion, New York, USA) following the manufacturer's protocol. Genome-wide gene expression data was obtained by hybridizing complementary RNA to Illumina's HumanHT-12v3 array and subsequently scanning these chips on the Illumina BeadArray Reader.

Replication blood dataset of 891 samples

We used a dataset comprising peripheral blood samples of 891 unrelated individuals from the Estonian Genome Centre, University of Tartu (EGCUT) biobank cohort of 53,000 samples for replication. Genotyping of these samples was performed according to Illumina's standard protocols, using Illumina Human370CNV arrays (Illumina Inc., San Diego, US), and imputed using IMPUTE v2[31], using the HapMap CEU phase 2 genotypes (release #24, build 36). Whole peripheral blood RNA samples were collected using Tempus Blood RNA Tubes (Life Technologies, NY, USA), and RNA was extracted using Tempus Spin RNA Isolation Kit (Life Technologies, NY, USA). Quality was measured by NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, DE, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). Whole-Genome gene-expression levels were obtained by Illumina Human HT12v3 arrays (Illumina Inc, San Diego, US) according to manufacturers' protocols.

Four non-blood primary tissues

Previously we described tissue-dependent eQTLs in 74 liver samples, 62 muscle samples, 83 SAT samples and 77 VAT samples from a cohort of 85 unrelated, obese Dutch individuals (all four tissues were available for 48 individuals)[18] (Table S5). These samples were genotyped according to standard protocols from Illumina, using Illumina HumanOmni-Quad BeadChips (Omni1). Genome-wide gene expression data of all samples was assayed by hybridizing complementary RNA to the Illumina HumanHT-12v3 array and then scanning it on the BeadArray Reader.

Cis-eQTL mapping

The method for normalization and principal component analysis-based correction of expression data, along with the methods to control population stratification and SNP quality, were described previously[17,18]. The *cis*-eQTL analysis was performed on probe-SNP combinations for which the distance from the center of the probe to the genomic location of the SNP was ≤ 250 kb. Associations were tested by non-parametric Spearman's rank correlation test and the P values were corrected for multiple testing by false-discovery rate (FDR) at $P < 0.05$,

30 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–2 (2010).

31 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5, e1000529 (2009).

in which the distribution was obtained from permuting expression phenotypes relative to genotypes 100 times within the HT12v3 dataset and comparing those with the observed P-value distribution. At FDR = 0.05 level, the P-value threshold was 2.4×10^{-4} for significantly associated probe-SNP pairs in blood, 1.5×10^{-5} in SAT, 5.21×10^{-6} in VAT, 6.3×10^{-6} in liver and 1.8×10^{-6} in muscle.

LincRNA function prediction

To predict the function(s) for lincRNAs, we interrogated the GeneNetwork database (www.genenetwork.nl/genenetwork) that has been developed in our lab (manuscript in preparation). In short, this database contains data extracted from approximately 80,000 microarray experiments that is publically available from the Gene Expression Omnibus; after extensive quality control, it contains data on 54,736 human, 17,081 mouse and 6,023 rat Affymetrix array experiments. Principal component analysis was performed on probe-set correlation matrices of each of four platforms (two human platforms, one mouse and one rat platform), resulting in 777, 377, 677 and 375 robust principal components, respectively. Jointly these components explain between 79% and 90% of the variance in the data, depending on the species or platform. Many of these components are well conserved across species and enriched for known biological phenomena. Because of this, we were able to combine the results into a multi-species gene network with 19,997 unique human genes, allowing us to utilize the principal components to accurately predict gene function by using a 'guilt-by-association' procedure (a description of the method is available at www.genenetwork.nl/genenetwork). Predictions were made based on pathways and gene sets from Gene Ontology, KEGG, BioCarta, TransFac and Reactome.

Functional annotation of lincRNA *cis*-eQTLs

We employed the HaploReg web tool26 to intersect SNPs (and their perfect proxies, $r^2 = 1$ using the CEU samples from the 1000 Genomes project) with regulatory information and also to calculate the fold enrichment of cell-type specific enhancers. In order to ascertain whether this enrichment was higher than expected, we took eQTL results from 100 permutations (shuffling the gene expression identifier labels): for each permutation we determined the top 112 eQTL probes and took the corresponding top SNPs and their perfect proxies ($r^2 = 1$). We extracted the fold enrichment of enhancers from HaploReg for these 100 sets of SNPs as well, which then permitted us to estimate the significance of enrichment of the real eQTL analysis, determined by fitting a normal distribution on the 100 log-transformed permutation enrichment scores.

Acknowledgments

The authors would like to thank Jackie Senior for editing the final text. We acknowledge EGCUT technical personnel, especially Mr. V. Soo and S. Smit. EGCUT data analyzes were carried out in part in the High Performance Computing Center of the University of Tartu.

Author Contributions

Conceived and designed the experiments: V.K., L.F. and CW.

Performed the experiments: V.K., H-J.W., J.K., D.V.Z., T.E., B.H., R.A., A.Z., E.R., U.V. and J.F.

Analyzed the data: V.K., H-J.W., J.K. and D.V.Z.

Contributed reagents/materials/analysis tools: M.H.H., R.S.N.F., A.M. and C.W.,

Wrote the paper: V.K., H-J.W., S.W., L.F. and C.W.

Supplemental Information

Figure S1
LincRNA probes show different expression characteristics compared to other transcripts. The figure shows the difference in quantile-normalized average expression intensity between lincRNA probes and non-lincRNA probes. The significance of difference in expression intensity was tested by the Wilcoxon Mann Whitney test.

Figure S2
Replicated lincRNA *cis*-eQTLs show identical allelic direction of effect in the both the discovery and replication datasets.
We compared the z-scores (association strength) of each significantly associated probe-SNP pair in the discovery dataset (Groningen HT12v3; N = 1,240) with the replication dataset (EGCUT; N = 891).

Figure S3
lincRNA probes with *cis*-eQTL effect show higher expression levels compared to lincRNA probes without *cis*-eQTL effect.
The significance of difference in expression intensity was tested by the Wilcoxon Mann Whitney test.

Figure S4
LincRNA *cis*-eQTL SNPs mostly affect lincRNA transcripts alone. Quantile-normalized average expression intensity of *cis*-eQTL lincRNAs and their neighboring protein coding genes without *cis*-eQTL.

Figure S5
Distribution of Z-scores of co-regulated lincRNA and protein-coding genes. We compared the z-scores (association strength) of each significantly associated probe-SNP pair for the 29 *cis*-eQTLs that affect both lincRNAs and protein-coding genes.

Figure S6
Number of specific and overlapping *cis*-eQTL lincRNAs identified across five different tissues.

Figure S7
Plots to show the association of age-related macular degeneration SNP rs13278062 with expression levels of lincRNA *LOC389641* and

protein-coding gene *TNFRSF10A* in blood (N = 1,249). The x-axis shows the number of samples according to the genotypes at rs13278062 and the y-axis is the average expression intensity of probes.

Figure S8
UCSC genome browser screen shot (http://genome.ucsc.edu) to show the location of age-related macular degeneration SNP, rs13278062. The x-axis is the chromosome location in the hg19 build and indicates the location of transcripts and regulatory elements identified by ENCODE on chromosome 8.

Figure S9
A plot to show the number of lincRNA *cis*-eQTLs on the y-axis within different regulatory regions on the x-axis.

Figure S10
Plots to show the *cis*-eQTL effect on lincRNA *XLOC_00197* from both microarray data (Groningen HT12v3; N = 1,240) and RNA-sequencing data (Montgomery *et al.*; N = 60). The x-axis shows the number of samples according to the genotypes at rs1120042 and rs2279692 (LD between these two SNPs, R2 = 0.96) in microarray data and RNA-sequencing data, respectively.

Table S1
LincRNA *cis*-eQTLs in blood and four other non-blood tissues.

Table S2
Function prediction of lincRNAs affected by *cis*-eQTLs using GeneNetwork.

Table S3
Identification of co-expressed genes for lincRNA *LOC389641* using GeneNetwork.

Table S4
Identification of co-expressed genes for lincRNA *LOC100131138* using GeneNetwork.

Table S5
Characteristics of sample cohorts used for *cis*-eQTL mapping.

DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts

PLoS Genetics, 2013 June; 9, e100359

Daria V. Zhernakova[1], Eleonora de Klerk[2], Harm-Jan Westra[1], Anastasios Mastrokolias[2], Shoaib Amini[2], Yavuz Ariyurek[2,3], Rick Jansen[4], Brenda W. Penninx[4], Jouke J. Hottenga[5], Gonneke Willemsen[5], Eco J. de Geus[5], Dorret I. Boomsma[5], Jan H. Veldink[6], Leonard H. van den Berg[6], Cisca Wijmenga[1], Johan T. den Dunnen[2,3], Gert-Jan B. van Ommen[2], Peter A. C. 't Hoen[2¤], Lude Franke[1¤*]

1 University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands
2 Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands
3 Leiden Genome Technology Center, Leiden, The Netherlands
4 Department of Psychiatry, The Netherlands Study of Depression and Anxiety, VU University Medical Center, Amsterdam, The Netherlands
5 Department of Biological Psychology, Netherlands Twin Registry, VU University, Amsterdam, The Netherlands
6 Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Utrecht, The Netherlands

Abstract

Many disease-associated variants affect gene expression levels (expression quantitative trait loci, eQTLs) and expression profiling using next generation sequencing (NGS) technology is a powerful way to detect these eQTLs. We analyzed 94 total blood samples from healthy volunteers with DeepSAGE to gain specific insight into how genetic variants affect the expression of genes and lengths of 3'-untranslated regions (3'-UTRs). We detected previously unknown *cis*-eQTL effects for GWAS hits in disease- and physiology-associated traits. Apart from *cis*-eQTLs that are typically easily identifiable using microarrays or RNA-sequencing, DeepSAGE also revealed many *cis*-eQTLs for antisense and other non-coding transcripts, often in genomic regions containing retrotransposon-derived elements. We also identified and confirmed SNPs that affect the usage of alternative polyadenylation sites, thereby potentially influencing the stability of messenger RNAs (mRNA). We then combined the power of RNA-sequencing with DeepSAGE by performing a meta-analysis of three datasets, leading to the identification of many more *cis*-eQTLs. Our results indicate that DeepSAGE data is useful for eQTL mapping of known and unknown transcripts, and for identifying SNPs that affect alternative polyadenylation. Because of the inherent differences between DeepSAGE and RNA-sequencing, our complementary, integrative approach leads to greater insight into the molecular consequences of many disease-associated variants.

Author Summary

Many genetic variants that are associated with diseases also affect gene expression levels. We used a next generation sequencing approach targeting 3' transcript ends (DeepSAGE) to gain specific insight into how genetic variants affect the expression of genes and the usage and length of 3'-untranslated regions. We detected many associations for antisense and other non-coding transcripts, often in genomic regions containing retrotransposon-derived elements. Some of these variants are also associated with disease. We also identified and confirmed variants that affect the usage of alternative polyadenylation sites, thereby potentially influencing the stability of mRNAs. We conclude that DeepSAGE is useful for detecting eQTL effects on both known and unknown transcripts, and for identifying variants that affect alternative polyadenylation.

Introduction

Genome-wide association studies (GWAS) have associated genetic variants, such as single nucleotide polymorphisms (SNPs) and copy number variants (CNVs), with numerous diseases and complex traits. However, the mechanisms through which genetic variants affect disease phenotypes or physical traits often remain unclear. To gain insight into these mechanisms, we have combined genotype data with gene expression data by conducting expression quantitative trait locus (eQTL) mapping. Previously, the level of gene expression was primarily assessed using oligonucleotide microarrays, which was a powerful method to profile the

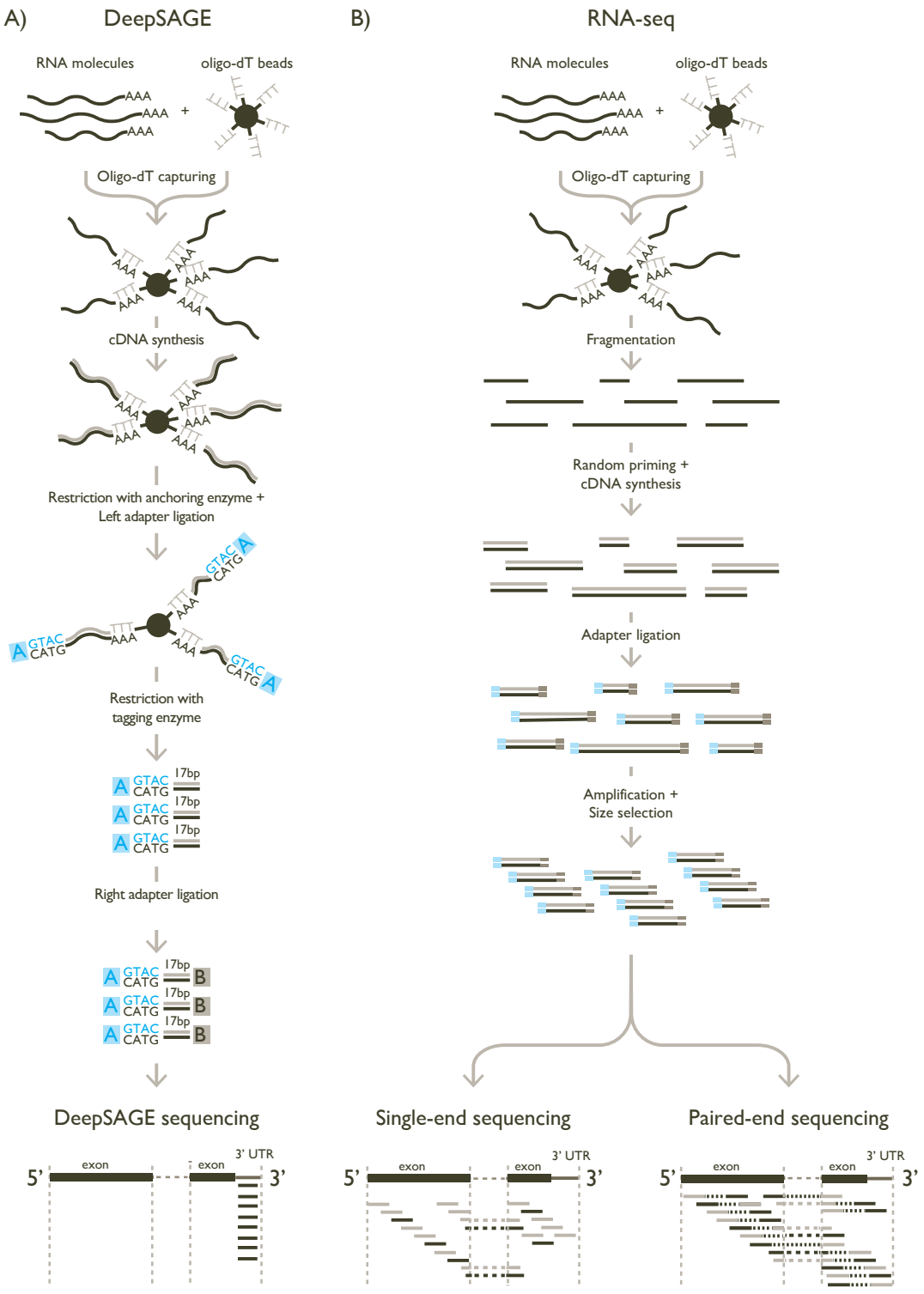

A)
DeepSAGE
RNA molecules
oligo-dT beads
AAA
Oligo-dT capturing
cDNA synthesis
Restriction with anchoring enzyme +
Left adapter ligation
A GTAC
CATG
Restriction with
tagging enzyme
17bp
Right adapter ligation
B
B)
RNA-seq
Fragmentation
Random priming +
cDNA synthesis
Adapter ligation
Amplification +
Size selection
DeepSAGE sequencing
Single-end sequencing
Paired-end sequencing
5'
3'
exon
3' UTR

Figure 1. Comparison of typical DeepSAGE and RNA-seq data generation steps.

A) DeepSAGE data preparation consists of the following basic steps: after RNA extraction the polyadenylated mRNA fraction is captured with oligo-dT beads. While RNA is still bound to the beads, double- stranded cDNA synthesis is performed. Next, cDNA is digested by NlaIII restriction enzyme (an anchoring enzyme), which cuts the DNA at CATG recognition sequences, leaving only the fragment with the most distal (3') CATG site associated with the beads. Subsequently, a GEX adapter is attached to the 5' end. This adapter contains a recognition sequence for the MmeI restriction enzyme that cuts the sequence 17 bp downstream of the CATG site. After ligation of a second GEX adapter, fragments containing 21 bp tags (17 unknown nucleotides + CATG) are ready for sequencing. B) A typical protocol for RNA-seq data preparation has the following steps: after RNA extraction the polyadenylated mRNA fraction is captured with oligo- dT beads. Captured RNA is fragmented and for each fragment cDNA synthesis is performed using random hexamer primers. Sequencing adapters are then ligated to each fragment. This is followed by size selection of the DNA fragments and PCR amplification. Then one end of the fragment is sequenced (single-end sequencing) or both ends (paired-end sequencing).

transcriptome[1–6]. But recently, high-throughput next generation sequencing (NGS) has become available, which allows quantification of expression levels by counting mRNA fragments (RNA-seq) or sequence tags (including serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE), and massively parallel signature sequencing (MPSS))[7].

To date, two NGS eQTL studies have been published[8,9], both of which used RNA-seq. Although RNA-seq is a versatile technique, the coverage in the ultimate 3'-end is usually lower due to the fragmentation and random hexamer priming steps involved in the sample preparation[10] (Figure 1B). DeepSAGE technology[11,12], however, concentrates on capturing information on the 3'-end of transcripts. In DeepSAGE, enzymatic cDNA digestions generate one specific tag of 17 nucleotides at the most 3'-CATG sequence of each transcript (Figure 1A). The majority of these 21-mer tags ('CATG' + 17 nucleotides) can be uniquely mapped to the genome to identify the genes expressed.

There are several features of NGS-based expression quantification methods that are especially important for eQTL mapping. While oligonucleotide arrays target a predefined set of transcripts or exons, both RNA-seq and DeepSAGE are capable of detecting novel and unannotated transcripts. If such a novel gene later turns out to be *cis*-regulated by trait- or disease-associated SNPs, it can represent an interesting causal candidate gene for the trait or disease under investigation. RNA-seq is extremely versatile, as it can quantify the expression of alternative transcripts, which makes it possible to detect SNPs regulating the choice between alternative transcripts. DeepSAGE, however, is generally not suited to detecting alternative splicing because of the 3' bias of the tag locations[13]. Because only sequence data is generated for these short tags, the read depth per tag is generally much greater than with RNA-seq, permitting accurate quantification of these tags[11,14]. Thus, this 3' emphasis makes DeepSAGE suitable for transcript variants that differ in 3'-UTRs and also for detecting alternative polyadenylation events, a widespread phenomenon that generates variation in 3'-UTR length[15,16]. Shortening or lengthening of the 3'-UTR may result in

1 Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–302 (2003).

2 Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437, 1365–9 (2005).

3 Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848–53 (2007).

4 Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* 452, 423–8 (2008).

5 Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302 (2010).

6 Fehrmann, R. S. N. *et al. Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* 7, e1002197 (2011).

7 Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98 (2011).

8 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–7 (2010).

9 Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–72 (2010).

10 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63 (2009).

11 't Hoen, P. A. C. *et al.* Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.* 36, e141 (2008).

the loss or gain of regulatory elements, such as miRNA binding sites or binding sites for proteins that can stabilize or destabilize the transcript[17,18]. Several SNPs that influence the choice for alternative polyadenylation sites have been detected by RNA-seq on a small number of individuals[19]. Here, we analyzed this phenomenon in more depth by performing *cis*-eQTL mapping on DeepSAGE data from total blood samples of 94 individuals.

Results

DeepSAGE dataset

For *cis*-eQTL mapping, we used DeepSAGE sequencing of 21 bp tags (16±7 million tags) from total blood samples from 94 healthy, unrelated individuals from the Netherlands Twin Register (NTR) and the Netherlands Study of Depression and Anxiety (NESDA)[20]. Sequence reads were mapped to the reference genome hg19 using Bowtie[21] and assigned to transcripts. We mapped 85±5% of tags to the genome and found that 77±9% of these mapped to exonic regions. Although 66±18% of these reads mapped to hemoglobin-alpha or –beta (*HBA1, HBA2, HBB*) genes, we were left with sufficient sequencing depth to detect a total of 9,562 genes at a threshold of at least two tags per million.

Cis-eQTL mapping

Once reads had been mapped, we quantified the expression levels of sequenced tags and performed *cis*-eQTL mapping, evaluating only those combinations of SNPs and tags that were located within a genomic distance of 250 kb, while using a Spearman rank correlation test (tag-level false discovery rate (FDR) controlled at 0.05). We identified 540 unique *cis*-regulated tags.

To subsequently increase the statistical power of eQTL detection, we used principal component analysis (PCA) to correct for technical and known and unknown biological confounders. The first principal components (PC) generally capture a high percentage of the expression variation, and these PCs mostly reflect technical, physiological and environmental variability.

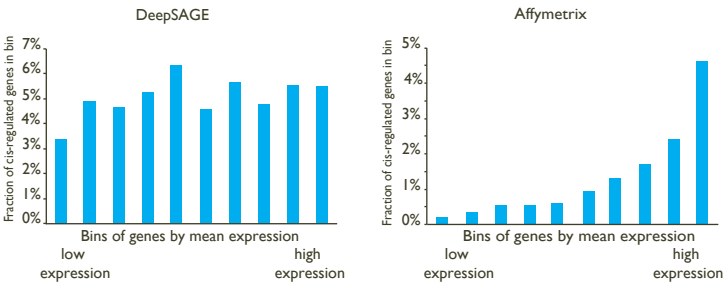


Figure 2. Fraction of *cis*-regulated genes in bins by mean gene expression levels for DeepSAGE and Affymetrix data.

For each dataset, all genes were sorted by their mean gene expression levels, and divided into ten equal bins. The X-axis reflects these bins, which are sorted by increasing mean gene expression levels. The Y-axis reflects the fraction of *cis*-regulated genes that fall into each bin.

12
Nielsen, K. L., Høgh, A. L. & Emmersen, J. DeepSAGE--digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Res.* 34, e133 (2006).

13
Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.* 20, 508–12 (2002).

14
Asmann, Y. W. *et al.* 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. BMC Genomics 10, 531 (2009).

15
Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* 33, 201–12 (2005).

16
Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* 22, 1173–83 (2012).

17
Barreau, C., Paillard, L. & Osborne, H. B. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res.* 33, 7138–50 (2005).

18
Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 320, 1643–7 (2008).

19
Yoon, O. K., Hsu, T. Y., Im, J. H. & Brem, R. B. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet.* 8, e1002882 (2012).

20
Penninx, B. W. J. H. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* 17, 121–40 (2008).

21
Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25 (2009).

Removing this variation allows for the detection of more eQTLs[6,22,23]. In our data the first principal component significantly correlated with sample GC content, and principal components 7 and 11 correlated with various blood cell count parameters (for details see Text S1, Figures S1 and S2). When using the PC corrected data, we observed an almost two-fold increase in the number of significant *cis*-eQTLs (1,011 unique *cis*-regulated tags, corresponding to 896 unique *cis*-regulated genes at tag-level FDR < 0.05). The list of detected eQTLs is given in Table S1.

Comparison with microarray results

We then compared the DeepSAGE *cis*-eQTLs with *cis*-eQTLs that we had identified using the Affymetrix HG-U219 expression microarrays on the same 94 samples. In that analysis we detected *cis*-eQTLs for only 274 genes (FDR < 0.05), only a third of what we identified using DeepSAGE.

We observed that this substantial difference could mostly be explained by the fact that the *cis*-eQTLs detected using Affymetrix microarrays nearly always reflected genes that are highly expressed in blood, whereas for DeepSAGE the detected *cis*-eQTL genes had expression levels that could be much lower (Figure 2). Although we only concentrated on tags that were expressed, there was no clear relationship between the mean tag level expression and the probability of showing a significant *cis*-eQTL. As such, DeepSAGE is much more capable of identifying *cis*-eQTLs for genes showing low expression than conventional microarrays.

It was therefore not a surprise that only 39% of the identified DeepSAGE *cis*-eQTLs could also be significantly detected in the microarray-based dataset (with identical allelic direction; Figure S3). Indeed, the *cis*-eQTLs that were not replicated in the microarray-based dataset generally had a much lower expression than the replicating *cis*-eQTLs (Wilcoxon Mann Whitney $P < 2 \times 10^{-3}$). And vice versa, we could significantly replicate 75% of the detected Affymetrix *cis*-eQTLs with the same allelic direction in the DeepSAGE data (Figure S3), indicating that DeepSAGE shows overlapping results with array-based data. At the same time, this provides insight into the regulation of gene expression by SNPs at many more loci.

We estimated the reduction that could be made in the sample size of the sequencing-based dataset to get the same number of *cis* regulated genes as in microarray-based data. We observed that the DeepSAGE sample size could be reduced by almost half (to 55 samples) to get the same number of significant *cis*-regulated genes as identified in the microarray analysis of the 94 samples. As such, these results clearly indicate that DeepSAGE has higher statistical power than microarrays.

Cis-eQTL effects on non-coding genes

While most microarray platforms interrogate mainly the protein-coding part of the transcriptome, NGS-based expression profiling will detect the majority of all expressed transcripts. Indeed, we detected eQTLs for known, but non-protein coding, genes: 8 antisense genes and 31 lincRNAs (Figure 3).

22
Biswas, S., Storey, J. D. & Akey, J. M. Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics* 9, 244 (2008).
23
Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–35 (2007).

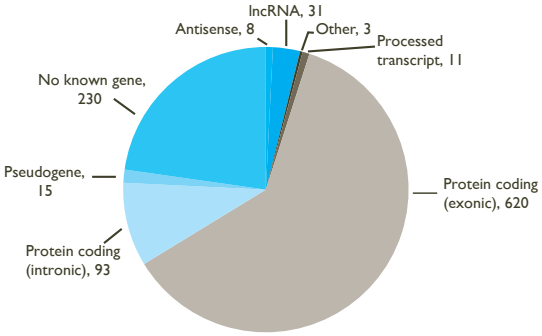


Figure 3. Mapping regions of *cis*-regulated tags.

The gene biotypes and exon/intron locations of unique *cis*-regulated tags, according to Ensembl v.69 annotation, are shown. The numbers indicate the number of tags mapping in the genes of the corresponding type.

We also expected to find a number of *cis*-eQTL effects on previously unknown transcripts. Of the 1,011 tags with a significant *cis*-eQTL effect, 230 did not map to known transcripts. Many of these tags map to retrotransposon-derived elements in the genome, which are known to be a source of novel exons[24]: 73 DeepSAGE tags with significant *cis*-eQTLs that did not map to annotated genes mapped to 72 unique LINE, SINE and LTR elements in the genome (Table 1).

New regulatory roles for disease- and trait-associated SNPs

We checked how many of our *cis*-acting SNPs were associated with complex traits or complex diseases ('trait-associated SNPs'), as published in the Catalog of Published Genome-Wide Association Studies. 104 of the 6,446 unique trait-associated SNPs were significant *cis*-eQTLs in our data (Table S2).

We were interested to determine whether the DeepSAGE data had revealed *cis*-eQTL effects for trait-associated SNPs that had been missed when using conventional arrays on much larger cohorts. We therefore compared our results to a re-analysis of a large-scale, array-based *cis*-eQTL mapping that we had conducted in whole peripheral blood samples when using a much larger sample size of 1,469 (using Illumina oligonucleotide arrays[6]).

Table 1.

Number of *cis*-regulated tags mapping to different genomic regions in tag-wise DeepSAGE eQTL mapping.

Type of genomic region	Number of *cis*-regulated tags
LINE	32
SINE	14
LTR	17

24 Cordaux, R. & Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691–703 (2009).

Table 2.

Trait-associated SNPs affecting DeepSAGE tags of 94 peripheral blood samples, but not detected in an array-based eQTL dataset of 1,469 peripheral blood samples.

SNP	Tag chr.	Tag position	In / near gene	Location in gene	Sense / antisense	Repeat masker	Associated trait
rs6704644	2	234380527	*DGKD*	3'-UTR	sense	None	Bilirubin levels
rs9875589	3	14196086	*XPC*	intron	antisense	LINE L1MB3	Ovarian reserve
rs4580814	5	1050754	*SLC12A7*	3'-UTR	sense	None	Hematological and biochemical traits
rs3194051	5	35884591	*IL7R*	-	noncoding	LINE L2C	Ulcerative colitis
rs4917014	7	50472441	*IKZF1*	3'-UTR	sense	None	Systemic lupus erythematosus
rs10092658	8	131017411	*FAM49B*	intron	sense	None	Protein quantitative trait loci
rs216345	9	33917317	*UBE2R2*	3'-UTR	sense	None	Bipolar disorder
rs12219125	10	20519590	*PLXDC2*	intron	sense	SINE AluJb	Diabetic retinopathy
rs7181230	15	40325714	*EIF2AK4*	intron	antisense	None	Dehydroepiandrosterone sulphate levels
rs4924410	15	40328035	*SRP14*	3'-UTR	sense	None	Ewing sarcoma
rs12594515	15	45995320	lincRNA *RP11-718O11.1*		noncoding	LTR MLT1A	Waist circumference, weight
rs6504218	17	62400467	*PECAM1*	3'-UTR	antisense	None	Coronary heart disease
	17	62397000	*PECAM1*	3'-UTR	antisense	LINE L1ME4A	Coronary heart disease
rs4072910	19	8640274	*MYO1F*	intron	sense	LINE MER1B	Height

We identified 13 trait-associated SNPs that did show a significant *cis*-eQTL effect in DeepSAGE eQTL mapping, but which did not show a *cis*-eQTL effect in the large, array-based, blood dataset (Table 2). This indicates that many trait-associated SNPs have regulatory effects that will, so far, likely have been missed using microarrays. While some of the tags map in the exons of annotated transcripts, we also found three *cis*-regulated tags in introns (sense direction), two tags antisense to the known transcripts, and two tags outside the annotated transcripts. These results indicate that several trait-associated SNPs affect the expression of previously unknown transcripts, adding functional relevance to SNPs and transcripts that are so far without annotation.

Some newly discovered eQTLs provide novel insights into genome-wide association hits for diseases or physiological traits, e.g. SNP rs216345, which has been associated with bipolar disorder. While it is located just downstream of *PRSS3*, we now saw that it also affects the expression of *UBE2R2*. There are many links between the ubiquitin system and bipolar disorder reported in the literature[25,26], making *UBE2R2* a more plausible candidate gene for bipolar disorder than *PRSS3*.

Genes with multiple SAGE tags and opposite allelic direction

In DeepSAGE, 21-bp-long cDNA fragments begin at the 'CATG' closest to the polyadenylation site (Figure 1). These individual 'tags' represent transcripts sharing the same polyadenylation site. If a SNP increases the abundance of one tag of a gene and decreases the abundance of another tag of the same gene, this indicates that the SNP is acting like a switch between transcripts with different 3'-UTRs or between alternative polyadenylation sites[19] (Figure 4).

25 Ryan, M. M. *et al.* Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Mol. Psychiatry* 11, 965–78 (2006).

26 Bousman, C. A. *et al.* Preliminary evidence of ubiquitin proteasome system dysregulation in schizophrenia and bipolar disorder: convergent pathway analysis findings from two independent samples. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* 153B, 494–502 (2010).

Table 3.

Cis-regulating SNPs significantly affecting multiple tags of the same gene in opposite directions.

Only significant eQTLs with FDR,0.05 for both *cis*-regulated tags were used.

SNP Name	SNP Type	Allele assessed	Probe chr.	Probe center	Overall Z-score	HGNC name	Annotation
rs12568757	G/A	G	1	150782318	4.404	*ARNT*	Alternative polyadenylation
			1	150782604	-4.314		
rs12566232	A/C	C	1	151846229	-6.859	*THEM4*	Alternative polyadenylation
			1	151846628	4.292		
rs1062826	G/C	C	1	160965239	-4.46	*F11R*	Alternative polyadenylation
			1	160966976	8.012		
rs13160562	G/A	A	5	96110323	-7.883	*ERAP1*	Alternative polyadenylation
			5	96111908	5.555		
rs3185733	A/C	A	5	112320282	4.027	*DCP2*	Alternative polyadenylation
			5	112356357	-4.46		
rs6948928	T/C	T	7	128589824	8.451	*IRF5*	Alternative polyadenylation
			7	128589265	-7.31		
rs2111903	G/C	C	12	47603121	5.143	*RP11-493L12.2*	Different exon
			12	47599911	-4.676		
rs841718	A/G	G	12	57489368	5.506	*STAT6*	Alternative polyadenylation
			12	57489809	-7.002		
rs2285934	G/T	T	12	113357275	4.931	*OAS1*	Different exon
			12	113355465	-5.191		
rs168822	C/T	T	16	55616984	7.37	*LPCAT2*	Alternative polyadenylation
			16	55620233	-4.664		
rs922446	T/C	T	16	56395733	-4.966	*AMFR*	Alternative polyadenylation
			16	56396100	4.392		
rs1674159	C/T	T	19	5915589	-7.103	*CAPS*	Alternative polyadenylation
			19	5916143	6.126		

Twelve genes with highly significant *cis*-eQTLs (p-value < 10^{-7}) contained tags that were regulated in opposite directions (Table 3). Most of the tags regulated in opposite direction could be explained by switches in alternative polyadenylation sites, as the tags were observed in the same last exon. The effect on alternative polyadenylation in *IRF5* has been found before[19,27] and was also validated in our cohort by RT-qPCR with primers in the proximal and distal parts of the 3'-UTR (Figure 5). As a further confirmation of the observed switches in using polyadenylation sites, we tested genotype-dependent alternative polyadenylation in two other RNA-seq datasets[8,9]. In these datasets, we confirmed the effect of two *cis*-regulating SNPs on *THEM4* and *F11R*. However, we could not confirm the effect of other SNPs on targets validated experimentally, including *IRF5*. This shows the limitation of RNA-seq data in detecting alternative polyadenylation events, likely due to limited and unequal coverage of the 3'-UTR. For only two genes, *OAS1* (also reported earlier[28]) and *RP11-493L12.2*, the observed opposite allelic effect originated from transcripts with different last exons, likely due to alternative splicing.

As we had identified several SNPs that affect alternative polyadenylation, we subsequently used a more permissive strategy, which required that, for a given SNP, only one eQTL tag should pass the FDR < 0.05 significance threshold while the other tag could be less significant. However, for such SNP-tag pairs, we then

27 Cunninghame Graham, D. S. *et al.* Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation. *Hum. Mol. Genet.* 16, 579–91 (2007).

28 Heap, G.A. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* 2, 1 (2009).

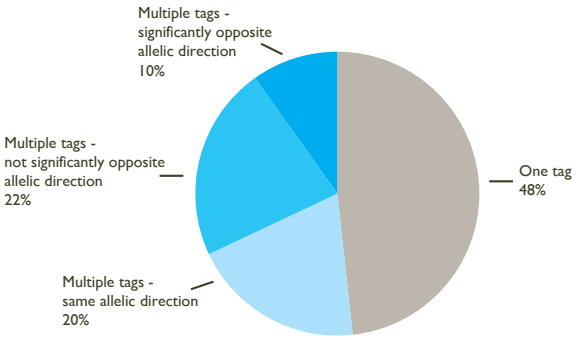


Figure 4. The number of *cis*-regulated tags per gene.

The percentages of *cis*-regulated tags mapping into the same gene are indicated (781 genes overall). For nearly half of the genes (48%) only one tag shows an eQTL effect. If multiple tags map within the same gene, only one eQTL tag should pass the FDR < 0.05 significance threshold while the other tag could be less significant. For these eQTLs the allelic direction is shown: same allelic direction (multiple tags within the same gene are *cis*-regulated by a SNP in the same direction), significantly opposite allelic direction (multiple tags within the same gene are *cis*-regulated by a SNP but with opposite directions and the difference between the correlation coefficients is significant), or opposite allelic direction but not significant (if the difference between correlation coefficients is not significant).

tested whether the allelic directions were opposite and if the difference between correlation coefficients was significant. With a differential correlation significance p-value threshold of 10^{-7}, we detected 41 unique genes showing regulation in opposite directions (Table S3). Of these, 23 (56%) showed opposite regulation of two tags in the same annotated 3'-UTR and a further 7 genes (17%) showed opposite regulation of tags in the same exons, both indicative of a switch in polyadenylation sites. Of these we picked *HPS1*, and validated a genotype-determined switch in preferred polyadenylation site usage by RT-qPCR analysis (Figure 5), indicating that the more permissive list also holds genuine changes in polyadenylation sites. The remaining 11 genes showed significant genotype-determined switches in expression of alternative transcripts not sharing the final exon. Thus, switches between shorter and longer 3'-UTRs occur more frequently than switches between transcripts with different 3'-UTRs.

To check whether such results appeared by chance, we took an equal number of top hits from a permuted eQTL run (shuffling the phenotype labels of the expression data, thus breaking the relationship between genotype and expression, but retaining linkage disequilibrium (LD) structure and structure in the expression data) and performed the same analysis as above (assessing an equal number of top eQTLs from the permuted analysis as we had investigated in the real analysis). Using the differential correlation significance threshold of 10^{-7} and conducting this permutation analysis ten times, we did not find any SNP that affected two tags in the same gene in a significantly different way, indicating this method is robust.

Since the eQTL SNPs are usually in strong LD with multiple SNPs, it is difficult to conclude whether a SNP is causal or which SNP is the likely causal variant. To identify the likely causal variant,

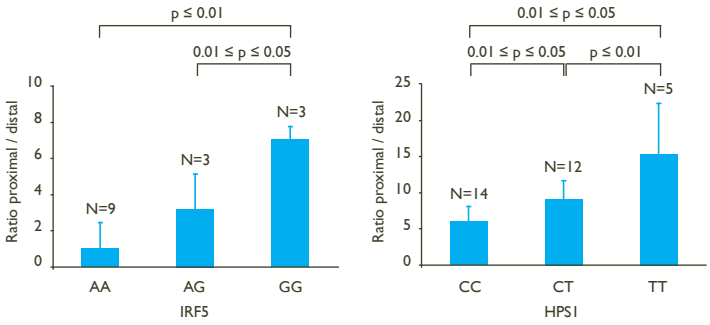


Figure 5. The choice of proximal/distal polyadenylation site in genes *IRF5* and *HPS1* depends on the genotypes of rs10488630 and rs11189600, respectively.

The ratio between the abundance of transcripts with proximal and distal 3'-UTR RT-qPCR products in *IRF5* (left) and *HPS1* (right) depends on the genotypes of *cis*-regulating SNPs rs10488630 and rs11189600, respectively. N denotes the number of individuals included in the analysis. These results indicate allele-specific preference for use of a proximal and distal polyadenlyation site.

we assessed whether any of these SNPs caused changes in poly-adenylation site usage. A direct effect on alternative poly-adenylation can be explained by a change in the polyadenylation site (corresponding to the cleavage site) or in the polyadenylation signal (a six-nucleotide motif located between 10–30 bases upstream of the cleavage site).

We searched for likely causative SNPs in linkage disequilibrium with the polyA-QTL SNP ($R^2 \geq 0.8$). We did not find any strong evidence for SNPs influencing the cleavage site and focused on *cis*-regulating SNPs located within polyadenylation signals. Considering the length and the motif of canonical and non-canonical polyadenylation signals[15], we performed a motif analysis in the sequence surrounding each *cis*-regulating SNP. We identified five SNPs that likely affect polyadenylation because there was a change in the polyadenylation signal (Table 4). As previously shown, rs10954213 causes the formation of a stronger polyadenylation signal in *IRF5*. Similar changes from non-canonical to stronger, canonical polyadenylation signals were observed for rs1062827 in *F11R* and rs6598 in *GIMAP5*. Moreover, rs12934747 creates a new canonical AATAAA polyadenylation signal in *LPCAT2*. The presence of this alternative polyadenylation signal at the beginning of the 3'-UTR leads to a decrease in transcripts containing the full length 3'-UTR, as observed by DeepSAGE (Figure 6). An opposite effect is observed for rs7063 in the ultimate 3'-end of the *ERAP1* gene, where the SNP causes the disruption of the strong canonical motif, and results in the use of a more proximal polyadenylation signal. Unfortunately we were not able to identify likely causative SNPs for each of these eQTLs. This could have several reasons: we imposed strict thresholds ($R^2 \geq 0.8$) on the LD between the detected *cis*-eQTLs and the putative causative SNPs; by imputing to the 1000 genomes dataset we may have missed causative SNPs unique to the Dutch population; and the list of experimentally validated polyadenylation sites is not exhaustive, because their detection depends on the expression level and cell type analyzed.

Table 4.

SNPs that likely affect polyadenylation due to a change in the polyadenylation signal. * This SNP was reported in [27] and is validated by our data.

Cis-regulating SNP	Causal SNP	R^2	SNP Type	Gene	Reference sequence	Alternative polyA signal	Distance to polyA site (bp)	Effect on 3'-UTR length
Formation/activation of polyA signal								
rs6948928	rs10954213	0.76*	G/A	*IRF5*	AATGAA	AATAAA	15	Shortening
rs168822	rs12934747	0.87	C/T	*LPCAT2*	AACAAA	AATAAA	27	Shortening
rs1062826	rs1062827	0.99	G/A	*F11R*	AGTAAA	AATAAA	21	Shortening
rs759011	rs6598	1	G/A	*GIMAP5*	AATAGA	AATAAA	13	Shortening
Disruption of polyA signal								
rs13160562	rs7063	1	T/A	*ERAP1*	AATAAA	AAAAAA	23	Shortening

Seven of the SNPs affecting polyadenylation are reported in the GWAS catalog as associated with diseases (Table S3), including rs2188962 and rs12521868, which are associated with Crohn's disease. We found that these SNPs were associated with a switch in the polyadenylation site of *IRF1*. This may reinforce previous evidence that *IRF1* is the gene in the IBD[5] locus responsible for its association with Crohn's disease[29]. *IRF1* is a family member of the *IRF5* gene. Thus, in the family of interferon regulatory factors, we found two members with genetic regulation of alternative polyadenylation sites, likely explaining susceptibility for Crohn's disease and systemic lupus erythematosus, respectively.

Another example is rs3194051, located in the *IL7R* gene. This SNP was not found in the analysis described above since it affects the expression of a tag on the same strand, downstream of *IL7R* in a LINE element (Table 2). However, this tag may represent an alternative 3'-UTR for *IL7R*. The SNP is associated with ulcerative colitis and *IL7R* may be another example of a gene in the inflammatory response pathway demonstrating alternative poly-adenylation.

Meta-analysis with RNA-seq data

In order to increase the statistical power to detect associations of SNPs with gene expression, we performed a first-of-its-kind eQTL mapping meta-analysis, combining DeepSAGE data with two published RNA-seq datasets. We used paired-end sequencing of mRNA derived from lymphoblastoid cell lines from HapMap individuals of European origin[8] and 35 and 46 bp single-end sequencing of mRNA derived from lymphoblastoid cell lines from HapMap individuals of Yoruba origin[9]. Sequence reads were mapped to the reference genome hg19 using Tophat[30] and assigned to transcripts. A consistently high percentage of reads (86–87% of aligned reads) mapped to exonic regions (Table 5).

We first performed eQTL mapping separately in all three datasets (Table 6), summarizing expression on the transcript level to permit comparisons between the datasets. The numbers of *cis*-regulated genes detected in transcript-wise analysis was lower than in tag-wise analysis, possibly because we missed resolution on alternative splicing- and alternative polyadenylation events. Again, PC correction greatly improved the number of *cis*-eQTLs detected (Table 6).

29 Huff, C. D. *et al.* Crohn's disease and genetic hitchhiking at IBD5. *Mol. Biol. Evol.* 29, 101–11 (2012).

30 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–11 (2009).

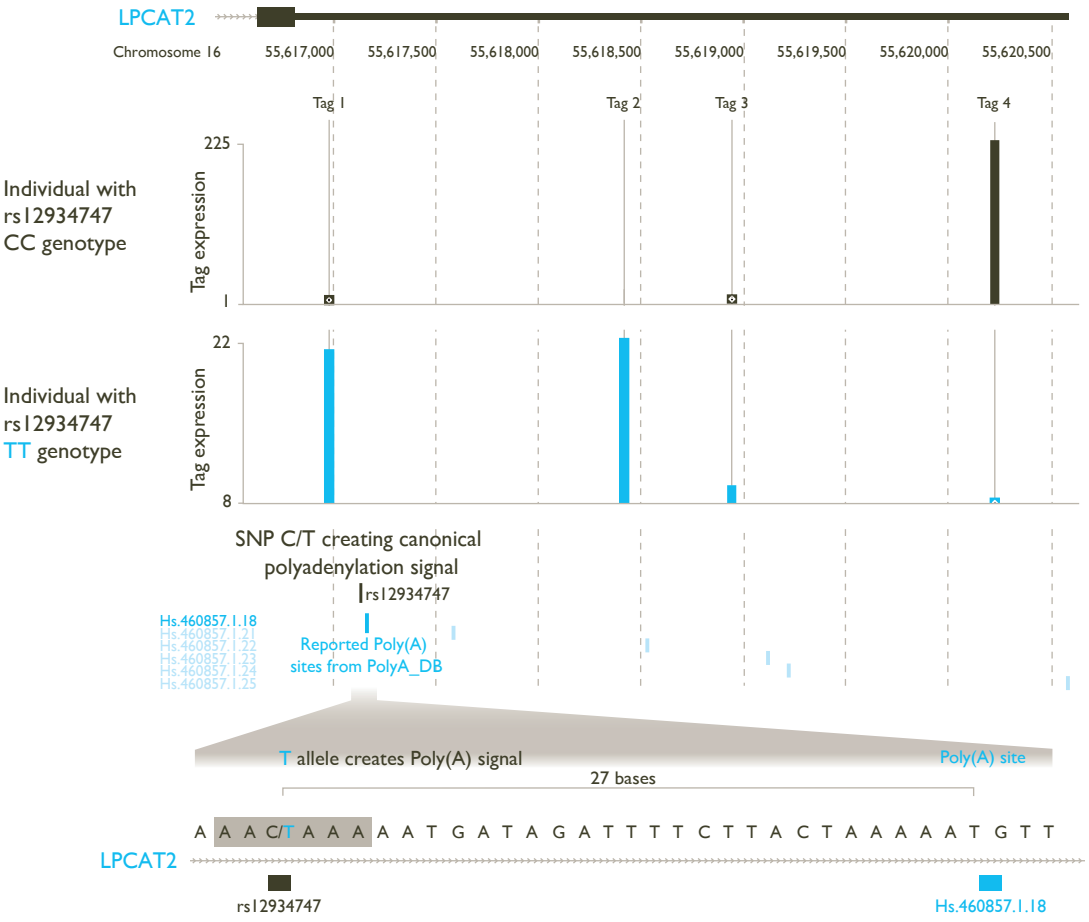


Figure 6. rs12934747*T creates a poly(A) signal in *LPCAT2* and leads to alternative polyadenylation site usage.

The y-axis represents the number of counts for the DeepSAGE tags. Two samples with different genotypes for SNP rs12934747, CC (reference allele) and TT (alternative allele), are shown as different traces. Below the coverage tracks, the position of rs12934747 is shown, together with the position of all reported polyadenylation sites from polyA_DB. An enlargement of the region containing the SNP is shown below. rs12934747 is located at the beginning of the 3-untranslated region (3'-UTR) of LPCAT2, 27 nucleotides upstream a reported and experimentally validated polyadenylation site. This SNP changes the sequence, creating a polyadenylation signal that leads to the usage of the reported polyadenylation site. The square block indicates the sequence of the polyadenylation signal created by rs12934747. The creation of this signal shortens the 3'-UTR, as indicated by the higher abundance of the proximal DeepSAGE just upstream of the polyadenylation signal, and the nearly absent distal DeepSAGE, in the sample with the TT genotype (both tags indicated by arrows). Tag 2 was filtered out because it was expressed in less than 90% of individuals. There is an additional tag 3 inbetween the proximal and distal tags, which is not *cis*-regulated.

Table 5.

Description of RNA next generation sequencing datasets.

Dataset		Sequencing type	Cell tissue type	Number of samples	Read length	Million reads per sample	Average % of mapped reads	Average % mapped reads in exons
Montgomery *et al.*		Paired-end RNA-seq	LCL	60	37 bp	9.5 ± 3	56	87
Pickrell *et al.*	Yale	Single-end RNA-Seq	LCL	72	35 bp	8.1 ± 2.3	85	86
Pickrell *et al.*	Argonne	Single-end RNA-Seq	LCL	72	46 bp	8.1 ± 1.8	80	86
NTR-NESDA		DeepSAGE	Total blood	94	21 bp	16 ± 7	85	88

We applied PC correction to the individual datasets. As for the DeepSAGE analysis, the first PC correlated strongly with the mean GC-percentage in the two RNA-seq datasets (Figure S1). We then assessed the robustness of the identified *cis*-eQTLs: we checked whether those in one dataset could be significantly replicated in the other two datasets. We observed that in each of the RNA-seq datasets approximately one-third of *cis*-eQTLs could be replicated in the other dataset (Table S4). The overlap between RNA-seq and DeepSAGE was smaller, reflecting differences in the two technologies, in cell types and in populations. In each comparison, we observed a very high concordance in the allelic direction of *cis*-eQTLs that could be replicated in another dataset.

We also looked at the replication of RNA-seq eQTLs in corresponding micro-array-based datasets. 80–88% of such eQTLs could be replicated in microarray data (Table S5). As we could cross-replicate many *cis*-eQTLs, we decided to conduct a meta-analysis to increase the statistical power. We calculated joint p-values using a weighted Z-score method. The number of *cis*-regulated genes then increased to 1,207 unique genes (Table 6; a list of detected eQTLs is given in Table S6), indicating that a meta-analysis of different types of sequencing based eQTL datasets reveals many more *cis*-regulated genes than the individual analyses.

For our meta-analysis results we determined the number of disease- and trait-associated SNPs using the Catalog of Published Genome-Wide Association Studies in the same way as for the DeepSAGE dataset. 107 of the 6,446 unique trait-associated SNPs

Table 6.

Number of detected *cis*-eQTLs in transcript-wise analysis of three harmonized RNA NGS datasets.

	Without principal component correction	With principal component correction
Montgomery *et al.* (paired-end RNA-seq)	94	145
Pickrell *et al.* (single-end RNA-seq)	199	438
NTR-NESDA transcript-wise (DeepSAGE)	292	579
Meta-analysis	651	1,207

showed a significant *cis*-eQTL effect in the meta-analysis. The overlap with 104 trait-associated SNPs detected in tag-wise DeepSAGE eQTL mapping was 37, indicating that the DeepSAGE transcript end quantification revealed other trait-associated *cis*-eQTLs than a meta-analysis on the level of whole transcripts. 21 of the 107 SNPs showed a significant *cis*-eQTL effect in the sequencing-based meta-analysis, but did not show a *cis*-eQTL effect in the large array-based blood dataset (Table 7).

Discussion

We have described the results from *cis*-eQTL mapping on DeepSAGE sequencing, a technique that is different from RNA-seq since it mainly targets the 3'-end of transcripts. We identified 1,011 unique *cis*-regulated tags (significant at tag-level FDR < 0.05). We performed eQTL mapping on the microarray expression data of the same samples and the number of detected *cis*-eQTLs was much smaller than in the DeepSAGE data, indicating the higher power of DeepSAGE in eQTL mapping. Moreover, for 220 of the *cis*-eQTLs SNPs detected by DeepSAGE we did not detect a significant *cis*-eQTL in a much larger microarray-based study in 1,469 whole peripheral blood samples[6]. 13 of these SNPs were reported as disease- or trait-associated in the GWAS catalog.

Table 7.

Trait-associated SNPs detected in the sequencing-based transcript-wise meta-analysis, but not detected in array-based eQTL dataset of 1,469 peripheral blood samples.

SNP name	Chr.	Transcript position (midpoint)	*Cis*-regulated gene	Associated trait
rs1052501	3	41963564	*ULK4*	Multiple myeloma
rs347685	3	141782879	*TFDP2*	Chronic kidney disease
rs4580814	5	1081324	*SLC12A7*	Hematological and biochemical traits
rs4947339	6	28911984	*C6orf100*	Platelet aggregation
rs2517532	6	31024818	*HCG22*	Hypothyroidism
rs2844665	6	31024818	*HCG22*	Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS-TEN)
rs6457327	6	31024818	*HCG22*	Follicular lymphoma
rs3130501	6	31324124	*HLA-B*	Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS-TEN)
rs2858870	6	32434437	*HLA-DRB9*	Nodular sclerosis Hodgkin lymphoma
rs3129889	6	32434437	*HLA-DRB9*	Multiple sclerosis
rs3135388	6	32434437	*HLA-DRB9*	Multiple sclerosis
rs477515	6	32434437	*HLA-DRB9*	Inflammatory bowel disease
rs9271100	6	32524134	*HLA-DRB6*	Systemic lupus erythematosus
rs9273349	6	32632106	*HLA-DQB1*	Asthma
rs3807989	7	116183034	*CAV1*	PR interval
rs12680655	8	135604552	*ZFAT*	Height
rs4929923	11	8642408	*TRIM66*	Menarche (age at onset)
rs12785878	11	71161461	*RP11-660L16.2*	Vitamin D insufficiency
rs12580100	12	56436876	*RPS26*	Psoriasis
rs4924410	15	40329664	*SRP14*	Ewing sarcoma
rs7364180	22	42184613	*MEI1*	Alzheimer's disease biomarkers

We observed that the number of *cis*-eQTLs detected in microarray data was higher in highly expressed genes, whereas for DeepSAGE the detected *cis*-eQTL genes had expression levels that could be much lower (Figure 2). This means that DeepSAGE is much better at identifying *cis*-eQTLs for genes showing low expression than conventional microarrays. This is because gene expression quantification using microarrays is more difficult as there is always a background signal present that needs to be accounted for. This is not the case for next-generation sequencing: although stochastic variation plays a major role in determining what RNA molecules will eventually be sequenced (especially for transcripts of low abundance), detection of such an RNA molecule is direct proof that it is being expressed.

Clearly, DeepSAGE can capture events that are likely to be missed by RNA-seq and conventional microarrays. It is not surprising, due to the different emphasis of DeepSAGE, that we could only replicate 39% of the DeepSAGE *cis*-eQTLs in the microarray data with a consistent allelic direction (Figure S3). The limited overlap between DeepSAGE- and microarray-based eQTL studies may be partly explained by the fixed thresholds applied, the interrogation of different transcript variants, and by the smaller dynamic range of microarrays. In addition, we found that more highly expressed genes were more often replicated than lower expressed ones.

Moreover, DeepSAGE allows for the detection of non-coding and novel transcripts not present on the microarrays. We showed that genetic variation affects the expression of a substantial number of lincRNAs and antisense genes, some of which have been linked to clinical traits. This suggests that clinical traits may be modified by expression of antisense transcripts or alternative 3'-UTR selection, which are not separately quantified in the microarray-based studies or in most RNA-seq, where standard protocols are still not strand-specific.

We also noticed a relatively high proportion of eQTLs with DeepSAGE tags mapping in SINE, LINE and LTR elements. These transposable elements contribute to the evolution and inter-individual variation of the human genome and to the diversification of the transcriptome, the latter facilitated by their inherent potential to be transcribed and the presence of cryptic splice acceptor and donor sites[24,31,32]. Some of the DeepSAGE tags we identified may be located in entirely new transcripts, but the majority is likely to represent alternative exons or 3'-UTRs of known transcripts, in accordance with the observed preferential location in introns or near genes.

Associations with transcripts and transcript variants not yet annotated may help to discover a function for these transcripts, as they are likely to play a role in the physiological and clinical traits associated with the SNP. Moreover, this will complement our knowledge of the pathways associated with these physiological and clinical traits.

In our study, we have demonstrated that genotype-dependent switches in the preference of alternative polyadenylation sites are common. One of these events has been well characterized: SNP

31
Belancio, V. P., Hedges, D. J. & Deininger, P. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res.* 34, 1512–21 (2006).
32
Kim, D.-S. *et al.* LINE FUSION GENES: a database of LINE expression in human genes. *BMC Genomics* 7, 139 (2006).

rs10954213 creates an alternative polyadenylation site in *IRF5*, shortens the 3'-UTR, stabilizes the mRNA, and increases *IRF5* expression, explaining its genetic association with systemic lupus erythematosus[19,27]. We have now discovered more examples where SNPs create or disrupt polyadenylation motifs. Amongst others, we identified a new, similar, genotype-dependent switch in preferred polyadenylation site for family member *IRF1*, with a probable link to Crohn's disease. Alternative polyadenylation associated with shortening of 3'-UTRs is a prominent event in the activation of immune cells[18]. Thus, genetically determined use of a proximal polyadenylation sites may predispose to inflammatory disorders such as Crohn's disease. The opposite correlations that we observed for most genes were slightly less pronounced than for *IRF5*. This indicates that mechanisms other than the creation or disruption of canonical polyadenylation motifs may also play a role. For example, SNPs in miRNA or protein-binding sites may specifically affect the stability of the transcript variant with the long 3'-UTR.

We subsequently conducted a *cis*-eQTL meta-analysis on the heterogeneous types of data using methods extended from those we developed for microarray-based eQTL meta-analysis[6]. We identified 1,207 unique *cis*-regulated genes. This number is substantially higher than in each of the datasets separately and indicates that different protocols for digital gene expression generally deliver consistent results. Nevertheless, the overlap at a fixed FDR of 0.05 is rather small, in particular between DeepSAGE and RNA-seq data. While this is partly attributable to using a strong threshold, there are other important reasons: firstly, the RNA-seq and DeepSAGE technologies frequently interrogate different transcript variants. Secondly, the RNA-seq studies were done on lymphoblastoid cell lines (LCLs) while the DeepSAGE study was on total blood, and some *cis*-eQTLs may be tissue-specific[33,34]. Finally, the DeepSAGE technology is strand-specific but the RNA-seq technologies evaluated here are not: where DeepSAGE will evaluate the expression of sense and antisense transcripts separately, RNA-seq will sum them. These reasons could explain why the percentage of RNA-seq-derived eQTLs that can be replicated by DeepSAGE is higher than the percentage of DeepSAGE-derived eQTLs that can be replicated by RNA-seq.

We conclude that DeepSAGE technology is useful to determine *cis*-eQTLs, as it is able to quantify the expression of novel transcripts, and to detect alternative polyadenylation effects and alternative 3'-UTR selection. It is complementary to other sequencing-based approaches, as they each reveal slightly different regulatory effects of genetic variants. Different sequencing-based eQTL analyses generally deliver consistent results, allowing for meta-analyses across different technologies. Future eQTL mapping studies based on DeepSAGE and other next generation sequencing strategies, using larger cohorts and different techniques, will likely reveal a more comprehensive picture of how far genetic variation affects the expression of protein-coding genes and non-coding RNAs.

33
Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* 8, e1002431 (2012).

34
Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* 7, e1002003 (2011).

Materials and Methods

Ethics statement

The medical ethical committee of the VUMC, Amsterdam, the Netherlands, approved the collection and analysis of material blood, DNA and RNA from the 94 participants from the Netherlands Twin Registry (NTR) and the Netherlands Study of Depression and Anxiety (NESDA).

NTR-NESDA dataset

We analyzed 21 bp DeepSAGE data from total blood RNA of 94 unrelated individuals who participated in NTR or NESDA. RNA was isolated using PaxGene tubes[20,35,36]. DeepSAGE sample preparation protocols, and alignment approaches were described in[37]. One sample was run on one lane of the Illumina GAII instrument. Data are available in ArrayExpress under accession number E-MTAB-1181.

The NTR-NESDA data was imputed using Beagle v3.1.0, with HapMap2 release 24 as a reference.

Tag mapping and expression estimation

Tags from DeepSAGE sequencing were aligned to the NCBI build 37 reference genome using Bowtie v. 0.12.7 allowing for a maximum of 1 mismatch and a maximum of 2 possible alignments (-n 1 -k 1 -m 2 --best --strata options).

The expression values were both quantified on an individual tag and transcript level. For the tag-wise analysis, the total number of occurrences of each unique tag in each sample was counted. We only included tags that were present in > 90% of samples. Tags with SNPs in the CATG recognition sequence (according to dbSNPv135) and the transcripts containing those tags were removed before eQTL analysis, since these SNPs can affect the position of the SAGE tag in the transcript. For the transcript-wise analysis, the tag counts for tags overlapping the exons of a transcript by at least half of the tag length were summed. Coordinates of LINE, SINE, LTR elements were derived from UCSC's RepeatMasker track (update: 2009-04-24).

GC content bias estimation

To calculate the GC content per individual for DeepSAGE data, GC frequencies for all mapped tags were summed after excluding the twenty most abundant tags, since their high abundance would give biased estimates.

Cis-eQTL mapping and correction for confounding effects through principal component analysis

Before eQTL mapping, transcript and tag expression values were quantile normalized. To perform *cis*-eQTL mapping, association of SNPs with the expression levels of tags or transcripts within a distance of 250 kb (as this is the average size of linkage regions) of the midpoint of the transcript or tag were tested with a non-parametric Spearman's rank correlation. Multiple testing correction was performed, controlling the false discovery rate (FDR) at 0.05. To determine the FDR, we created a null distribution by repeating the *cis*-eQTL analysis after permuting the sample labels 10 times[38].

35 Maugeri, N. *et al.* LPAR1 and ITGA4 regulate peripheral blood monocyte counts. *Hum. Mutat.* 32, 873–6 (2011).

36 Willemsen, G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. Twin Res. *Hum. Genet.* 13, 231–45 (2010).

37 Hestand, M. S. *et al.* Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. *Nucleic Acids Res.* 38, e165 (2010).

We argue that gene expression levels from NGS-based datasets are, like micro-array based data, derived from genetic, technical and environmental effects. As such, compensating for these non-genetic effects would increase the power to detect *cis*-eQTL effects. To mitigate the effects of non-genetic sources of variability, we first $\log_2$ transformed the data and centered and scaled each tag, and subsequently applied PCA on the sample correlation matrix. We then used the first PCs as covariates, and re-did the non parametric *cis*-eQTL mapping on the residual expression data (using the procedure described by[6]).

Validation of genotype-dependent alternative polyadenylation in RNA-seq datasets

The genomic coordinates of the 3'-UTR, obtained from Refseq Genes, were split into two separate regions (distal and proximal 3'-UTRs) according to the position of the DeepSAGE tags with opposite directions, the position of LongSAGE tags from CGAP, and the position of reported and predicted polyadenylation sites from polyA_DB database. To calculate the coverage in proximal and distal regions in RNA-seq datasets, we created a coverage histogram from each .bam alignment file using coverageBed tool from BEDTools package (version 2.17.0)[39]. Subsequently, a custom Python script was used to convert the histogram in number of nucleotides mapped per region, normalized by the length of the region. The ratio between the number of counts in the proximal region and the distal region was then calculated.

qPCR validation of alternative polyadenylation

Expression of short and long variants of *HPS1* and *IRF5* was quantified by qRT-PCR, which was performed on a subset of RNA samples used for the DeepSAGE sequencing. cDNA was synthesized from 400 ng of total RNA using BioScript MMLV Reverse Transcriptase (Bioline) with 40 ng of random hexamer and oligo(dT)18 primers following manufacturer's instructions (for the list of primer sequences see Table S7). Primers specific to short or long variants of *HPS1* were designed using Primer3Plus program, primers for *IRF5* were designed as previously described[40]. qRT-PCR was performed on the LightCycler 480 (Roche) using 26 SensiMix reagent (Bioline). 45 cycles of two-step PCR were performed for HPS1, and 55 cycles of three-step PCR (95°C for 15 s, 48°C for 15 s, and 60°C for 40 s) for *IRF5*. Each measurement was performed in duplicates. PCR efficiency was determined using the LinRegPCR program[41] v.11.1 according to the described method[42]. Ratios between distal and proximal PCR products were then calculated and significance was tested performing a T-test.

Identifying causal SNPs affecting polyadenylation

We obtained all the proxy SNPs for all SNPs identified as *cis*-regulating the choice of polyadenylation site. To do this we used bi-allelic SNPs that pass QC from the 1000G European panel (v3.20101123) and took all SNPs that were in linkage disequilibrium with the query SNPs ($R^2 \geq 0.8$, distance between SNPs within 250 kb).

From this list of *cis*-regulating SNPs in linkage disequilibrium, we kept only SNPs, which were located in the *cis*-regulated genes. The filtering was performed by intersecting .bed files containing

38
Breitling, R. *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 4, e1000232 (2008).
39
Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–2 (2010).
40
Graham, R. R. *et al.* Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl. Acad. Sci. U. S. A.* 104, 6758–63 (2007).
41
Ramakers, C., Ruijter, J. M., Deprez, R. H. L. & Moorman, A. F. M. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci. Lett.* 339, 62–6 (2003).
42
Ruijter, J. M. *et al.* Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.* 37, e45 (2009).

SNPs coordinates and coordinates of *cis*-regulated genes from the RefSeq database, using the table browser tool in the UCSC genome browser and the overlap intervals tool in Galaxy (version 1.0.0).

Intersection of SNPs with validated and predicted polyadenylation sites was performed using annotation in the PolyA-DB database (PolyA_DB 1 and PolyA_SVM) on UCSC (table browser tool). Detection of SNPs within polyadenylation signals was performed by extracting the strand specific sequence five nucleotide upstream and downstream each SNP (using table browser tool in UCSC) and performing a motif search using custom Perl script. Canonical and non-canonical polyA motifs searched were AATAAA, ATTAAA, TATAAA, AGTAAA, AAGAAA, AATATA, AATACA, CATAAA, GATAA, AATGAA, TTTAAA, ACTAAA, and AATAGA. For every SNP located in a putative polyadenylation signal motif, the distance to validated and predicted polyadenylation sites from PolyA-DB was calculated. Only motifs within a distance of 30 nucleotides from a polyadenylation site were considered true polyadenylation signals. Newly formed polyadenylation signals were detected by changing the reference allele of the SNP with the alternative allele, followed by the same polyadenylation signal motif search using custom Perl scripts.

For the *cis*-regulated genes where the SNP is located within a true polyadenylation signal, we retrieved the coverage of every SAGE tag upstream and downstream the putative affected polyadenylation site and calculated the ratio between proximal and distal tags for the different genotypes to confirm the expected effects of polyadenylation site formation or disruption.

RNA-seq datasets

For the meta-analysis we combined DeepSAGE data with two published RNA-seq datasets. The first dataset was 37 bp paired-end RNA-sequencing data from HapMap individuals[8] (ArrayExpress:E-MTAB-197): RNA from lymphoblastoid cell lines of 60 HapMap CEPH individuals was sequenced on the Illumina GAII sequencer, while genotype data had already been generated within the HapMap project.

The second dataset was single-end RNA-sequencing data from HapMap individuals[9,43] (GEO:GSE19480 and at http://eqtl.uchicago.edu/RNA_Seq_data/): RNA was sequenced from lymphoblastoid cell lines of 72 HapMap Yoruba individuals from Nigeria on the Illumina GAII platform in two sequencing centers: Yale (using 35 bp reads) and Argonne (using 46 bp reads).

Since the Montgomery *et al.* paper used genotype data for some individuals that were not in the HapMap3 panel (NA0851, NA12004, NA12414 and NA12717), we imputed these individuals using Beagle v3.1.0, with HapMap2 release 24 as a reference.

RNA-seq read mapping

Reads from single- and paired-end RNA-sequencing were mapped to the human genome NCBI build 37 (reference annotation from Ensembl GRCh37.65) using Tophat v. 1.3.330 – a splice-aware aligner that maps RNA-seq reads to the reference genome using Bowtie[21]. We used default settings (maximum 2

43 Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* 6, e1001236 (2010).

mismatches, 20 possible alignments per read) with a segment length value of 17 bp. Reads that corresponded to the flag 1796 in the .bam alignment file (read unmapped, not primary alignment, read fail quality check, read is PCR or optical duplicate) were filtered out. The numbers of raw and mapped reads for each dataset are given in Table 5.

Read quantification

To estimate expression levels in RNA-seq data, reads that overlapped with exons from known transcripts (GRCh37.65) were quantified using the coverageBed method from BEDTools suite[39].

For transcript level quantification the read count *C* for sample *s* for transcript *tr* was calculated as a sum of expression values over all exons contained in this transcript:

$$C_{s,tr} = 10^6 \times \sum (n_e * B_e)$$

where n_e is number of reads overlapping exon e by not less than half of read's length, and B_e is the breadth of coverage for exon e (% of exon length covered by the reads mapping to that exon).

In case a read mapped to multiple transcripts, the read was counted for all transcripts, since the short reads are difficult to assign to a specific transcript. Multiplication by breadth of coverage was performed to help in distinguishing between different isoforms by assigning higher weight to exons fully covered by reads in contrast to alternative exons covered only partly.

Because different methods have different capacity to identify alternative splicing events, we subsequently summarized our eQTL results to unique genes.

Meta-analysis

Meta-analysis was conducted by using a weighted Z-method, weighing each of the datasets by the square root of the number of samples per dataset[6].

Microarray datasets

We compared the results to corresponding microarray dataset eQTL mapping results. For each of the 94 individuals from NTR NESDA study, Affymetrix HG-U219 expression data were generated at the Rutgers University Cell and DNA Repository (RUCDR, http://www.rucdr.org). NTR and NESDA samples were randomly assigned to plates with seven plates containing subjects from both studies to better inform array QC and study comparability. Gene expression data were required to pass standard Affymetrix QC metrics (Affymetrix expression console) before further analysis. Probe sets were removed when their mapping location was ambiguous or if their location intersected a polymorphic SNP (dropped if the probe oligonucleotide sequence did not map uniquely to hg19 or if the probe contained a polymorphic SNP based on HapMap3[44] and 1000 Genomes[45] project data). Expression values were obtained using RMA normalization implemented in Affymetrix Power Tools (APT, v 1.12.0). *MixupMapper* revealed no sample mix-ups[46].

44 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–8 (2010).

45 Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–73 (2010).

46 Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* 27, 2104–2111 (2011).

For RNA-seq data we used corresponding microarray datasets that were available for most of the individuals present in RNA-seq datasets. We used Illumina expression data provided by Stranger *et al.*[3] of the 72 HapMap YRI individuals (56 of which were also present in RNA-seq dataset from Pickrell *et al.*) and 60 HapMap CEU individuals provided by Montgomery *et al.* (58 of which were also present in RNA-seq dataset from Montgomery *et al.*).

The same normalization procedure was performed as for the sequencing-based datasets: quantile normalization, and subsequent probe set centering to zero, z-score transformation, and scaling to a standard deviation of one.

Data access

The newly generated DeepSAGE data for NTR-NESDA dataset is available in ArrayExpress under accession number E MTAB-1181 (ENA: ERP001544).

Author Contributions

Conceived and designed the experiments: P.A.C.t.H., L.F., G.J.B.v.O. and J.T.d.D.

Performed the experiments: P.A.C.t.H., Y.A. and A.M.

Analyzed the data: D.V.Z., E.d.K., P.A.C.t.H., H.J.W. and S.A.

Contributed reagents/materials/analysis tools: P.A.C.t.H., H.J.W., R.J., B.W.P., J.J.H., G.W., E.J.d.G., D.I.B., J.H.V., L.H.v.d.B. and C.W.

Wrote the paper: D.V.Z., E.d.K., P.A.C.t.H. and L.F.

Supplementary Information

Figure S1
Correlation of GC content with principal component 1 (PC1) eigenvector coefficients for all the three datasets. Pearson correlation coefficient and corresponding p-value are shown in the plot.

Figure S2
Blood cell counts in DeepSAGE data captured by the eigenvector coefficients on principal components PC7 (left) and PC11 (right). Experimentally determined blood cell counts at the time of RNA isolation were available for 36/94 samples. Blood cell counts are expressed as (number of cells) x 10^9/L. Pearson correlation coefficients and corresponding p-values are shown in the plot.

Figure S3
Replication of Affymetrix eQTLs in DeepSAGE dataset and DeepSAGE eQTLs in Affymetrix data. The numbers of unique *cis*-regulated genes is given after each filtering step.

Table S1
List of detected eQTLs in tag-wise eQTL mapping.

Table S2
Trait-associated SNPs affecting the expression of DeepSAGE tags of 94 peripheral blood samples.

Table S3
List of candidate genes with alternative polyadenylation event detected using a permissive strategy.

Table S4
Replications between RNA-seq and DeepSAGE eQTLs.

Table S5
Replication of RNA-seq eQTLs in microarray-based datasets.

Table S6
List of detected eQTLs in the meta-analysis.

Table S7
Primer sequences for qPCR validation.

Text S1
Additional details on principal component analysis of DeepSAGE expression data.

Systematic identification of *trans*-eQTLs as putative drivers of known disease associations

Nature Genetics, 2013 October; 45: 1238–1243

Harm-Jan Westra[1,40], Marjolein J. Peters[2,3,40], Tõnu Esko[4,40], Hanieh Yaghootkar[5,40], Claudia Schurmann[6,40], Johannes Kettunen[7,8,40], Mark W. Christiansen[9,40], Benjamin P. Fairfax[10,11], Katharina Schramm[12,13], Joseph E. Powell[14,15], Alexandra Zhernakova[1], Daria V. Zhernakova[1], Jan H. Veldink[16], Leonard H. Van den Berg[16], Juha Karjalainen[1], Sebo Withoff[1], André G. Uitterlinden[2,3,17], Albert Hofman[3,17], Fernando Rivadeneira[2,3,17], Peter A. C. 't Hoen[18], Eva Reinmaa[4], Krista Fischer[4], Mari Nelis[4], Lili Milani[4], David Melzer[19], Luigi Ferrucci[20], Andrew B Singleton[21], Dena G. Hernandez[21,22], Michael A. Nalls[21], Georg Homuth[6], Matthias Nauck[23], Dörte Radke[24], Uwe Völker[6], Markus Perola[4,8], Veikko Salomaa[8], Jennifer Brody[9], Astrid Suchy-Dicey[25], Sina A. Gharib[26], Daniel A Enquobahrie[25], Thomas Lumley[27], Grant W Montgomery[28], Seiko Makino[10], Holger Prokisch[12,13], Christian Herder[29], Michael Roden[29–31], Harald Grallert[32], Thomas Meitinger[12,13,33,34], Konstantin Strauch[35,36], Yang Li[37], Ritsert C Jansen[37], Peter M. Visscher[14,15], Julian C Knight[10], Bruce M. Psaty[9,38,41], Samuli Ripatti[7,8,39,41], Alexander Teumer[6,41], Timothy M. Frayling[5,41], Andres Metspalu[4,41], Joyce B. J. van Meurs[2,3,41] & Lude Franke[1,41]

7

1. Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
2. Department of Internal Medicine, Erasmus Medical Centre Rotterdam, Rotterdam, The Netherlands
3. Netherlands Genomics Initiative-sponsored Netherlands Consortium for Healthy Aging (NGI-NCHA), Leiden and Rotterdam, The Netherlands
4. Estonian Genome Center, University of Tartu, Tartu, Estonia
5. Genetics of Complex Traits, University of Exeter Medical School, Exeter, UK
6. Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
7. Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
8. Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland
9. Cardiovascular Health Research Unit, University of Washington, Seattle, Washington, USA
10. Wellcome Trust Centre for Human Genetics, Oxford, UK
11. Department of Oncology, Cancer and Haematology Centre, Churchill Hospital, Oxford, UK
12. Institute of Human Genetics, Helmholz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany
13. Institute of Human Genetics, Technical University Munich, Munich, Germany
14. University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland, Australia
15. Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia
16. Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Utrecht, The Netherlands
17. Department of Epidemiology, Erasmus Medical Centre Rotterdam, Rotterdam, The Netherlands
18. Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands
19. Institute of Biomedical and Clinical Sciences, University of Exeter Medical School, Exeter, UK
20. Clinical Research Branch, National Institute on Aging–Advanced Studies in Translational Research on Aging (ASTRA) Unit, Harbor Hospital, Baltimore, Maryland, USA
21. Laboratory of Neurogenetics, National Institute on Aging, US National Institutes of Health, Bethesda, Maryland, USA
22. Department of Molecular Neuroscience, Reta Lila Laboratories, Institute of Neurology, University College London, London, UK
23. Institute for Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany
24. Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
25. Department of Epidemiology, University of Washington, Seattle, Washington, USA
26. Computational Medicine Core, Center for Lung Biology, Division of Pulmonary & Critical Care Medicine, Department of Medicine, University of Washington, Seattle, Washington, USA
27. Department of Statistics, University of Auckland, Auckland, New Zealand
28. Queensland Institute of Medical Research, Herston, Queensland, Australia
29. Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany
30. Department of Endocrinology and Diabetology, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
31. Department of Metabolic Diseases, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
32. Research Unit of Molecular Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany
33. German Center for Cardiovascular Research (DZHK), Göttingen, Germany
34. Munich Heart Alliance, Munich, Germany
35. Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians Universität, Neuherberg, Germany
36. Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany
37. Groningen Bioinformatics Center, University of Groningen, Groningen, The Netherlands
38. Group Health Research Institute, Group Health Cooperative, Seattle, Washington, USA
39. Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
40. These authors contributed equally to this work
41. These authors jointly directed this work.
Correspondence should be addressed to L. Franke (lude@ludesign.nl).

Identifying the downstream effects of disease-associated SNPs is challenging. To help overcome this problem, we performed expression quantitative trait locus (eQTL) meta-analysis in non-transformed peripheral blood samples from 5,311 individuals with replication in 2,775 individuals. We identified and replicated *trans*-eQTLs for 233 SNPs (reflecting 103 independent loci) that were previously associated with complex traits at genome-wide significance. Some of these SNPs affect multiple genes in *trans* that are known to be altered in individuals with disease: rs4917014, previously associated with systemic lupus erythematosus (SLE)[1], altered gene expression of *C1QB* and five type I interferon response genes, both hallmarks of SLE[2–4]. DeepSAGE RNA sequencing showed that rs4917014 strongly alters the 3'-UTR levels of *IKZF1* in *cis*, and chromatin immunoprecipitation and sequencing analysis of the *trans*-regulated genes implicated *IKZF1* as the causal gene. Variants associated with cholesterol metabolism and type I diabetes showed similar phenomena, indicating that large-scale eQTL mapping provides insight into the downstream effects of many trait-associated variants.

Genome-wide association studies (GWAS) have identified thousands of variants that are associated with complex traits and diseases. However, because most variants are noncoding, it is difficult to identify causal genes. Several eQTL-mapping studies[5–8] have shown that disease-predisposing variants often affect the gene expression levels of nearby genes (*cis*-eQTLs). A few recent studies have also identified *trans*-eQTLs[5,9–13], showing the downstream consequences of some variants. However, the total number of reported *trans*-eQTLs is low, mainly owing to the multiple-testing burden. To improve statistical power, we performed an eQTL meta-analysis in 5,311 peripheral blood samples from 7 studies (EGCUT[14], InCHIANTI[15],

1 Han, J.-W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* 41, 1234–7 (2009).

2 Bengtsson, A. A. *et al.* Activation of type I interferon system in systemic lupus erythematosus correlates with disease activity but not with antiretroviral antibodies. *Lupus* 9, 664–71 (2000).

3 Bohlson, S. S., Fraser, D. A. & Tenner, A. J. Complement proteins C1q and MBL are pattern recognition molecules that signal immediate and long-term protective immune functions. *Mol. Immunol.* 44, 33–43 (2007).

4 Ytterberg, S. R. & Schnitzer, T. J. Serum interferon levels in patients with systemic lupus erythematosus. *Arthritis Rheum.* 25, 401–6 (1982).

5 Fehrmann, R. S. N. *et al.* *Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* 7, e1002197 (2011).

6 Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6, e1000888 (2010).

7 Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–72 (2010).

8 Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302 (2010).

9 Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44, 502–10 (2012).

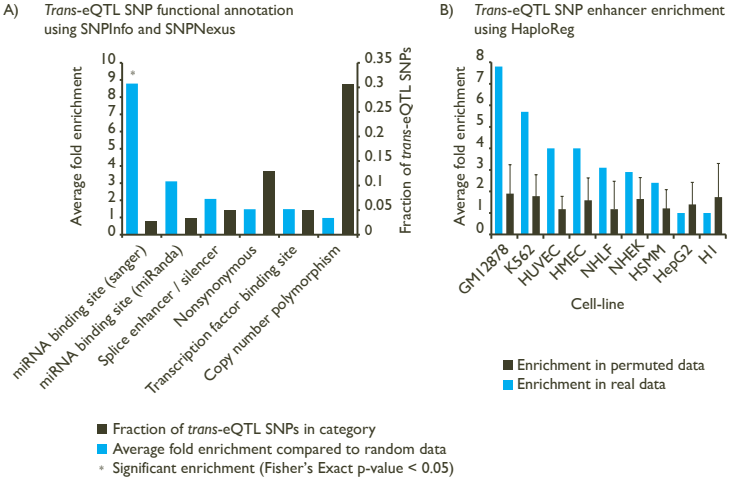


Figure 1.

Trans-eQTL SNPs are enriched for functional elements. We investigated whether *trans*-eQTL SNPs are enriched for certain functional elements using the online tools SNPInfo, SNPNexus and HaploReg that rely on data from, among others, the ENCODE Project.

A) *Trans*-eQTL SNPs are enriched for mapping within miRNA binding sites.

B) *Trans*-eQTL SNPs show strong enrichment (as annotated using HaploReg) for enhancer regions that are present in K562 (myeloid) and GM12878 (lymphoid) cell lines (error bars, 1 s.d.).

Table 1. Results of *cis*- and *trans*-eQTL mapping analyses

	***Cis*-eQTL analysis**		***Trans*-eQTL analysis**	
	FDR < 0.05 significance	Bonferroni significance	FDR < 0.05 significance	Bonferroni significance
Number of significant unique SNP-Probe pairs	664,097	395,543	1,513	643
Number of significant unique eQTL SNPs	397,310	266,036	346	200
Number of significant unique eQTL probes	8,228	5,738	494	240
Number of significant unique eQTL genes	6,418	4,690	430	223
Number of significant unique eQTL probes not mapping to genes	636	326	35	13

Rotterdam Study[16], Fehrmann[5], HVH[17–19], SHIP-TREND[20] and DILGOM[21]) and replication analysis in another 2,775 samples. We aimed to ascertain to what extent SNPs affect genes in *cis* and in *trans* and to determine whether eQTL mapping in peripheral blood could identify downstream pathways that might be drivers of disease processes.

Our genome-wide analysis identified *cis*-eQTLs for 44% of all tested genes (6,418 genes at probe-level false discovery rate (FDR) < 0.05 and 4,690 genes with a more stringent Bonferroni multiple-testing correction; Table 1, Supplementary Figures 1–3 and Supplementary Tables 1–3). Our *trans*-eQTL analysis focused on 4,542 SNPs that have been implicated in complex disease or traits (derived from the Catalog of Published GWAS; see URLs). In the discovery data set, we detected 1,513 significant *trans*-eQTLs that included 346 unique SNPs (FDR < 0.05; 8% of all tested SNPs; Table 1, Supplementary Figure 4 and Supplementary Table 4) affecting the expression of 430 different genes (643 *trans*-eQTLs, including 200 unique SNPs and 223 different genes with a more stringent Bonferroni correction).

We used stringent procedures for *trans*-eQTL detection (Supplementary Note) and various benchmarks to ensure reliability: for 26 *trans*-eQTL genes, the eQTL SNP affected multiple probes within these genes (Supplementary Table 5), always with consistent allelic directions, suggesting that our probe-filtering procedure was effective in preventing false-positive *trans*-eQTLs. *Trans*-eQTLs showed similar effect sizes across the various cohorts (Supplementary Figure 5).

We did not find evidence that *trans*-eQTLs were driven by differences in age or blood cell counts between individuals (Supplementary Figure 6, Supplementary Table 6 and Supplementary Note). However, we cannot exclude this possibility entirely because FACS analyses on individual cell types had not been conducted. We also detected previously reported blood *trans*-eQTLs[5] in this study (Supplementary Figure 7, Supplementary Table 7 and Supplementary Note).

10
Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* 7, e1002078 (2011).
11
Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* 44, 1084–9 (2012).
12
Heinig, M. *et al.* A *trans*-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* 467, 460–4 (2010).
13
Small, K. S. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat. Genet.* 43, 561–4 (2011).
14
Metspalu, A. The Estonian Genome Project. *Drug Dev. Res.* 62, 97–101 (2004).
15
Tanaka, T. *et al.* Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet.* 5, e1000338 (2009).
16
Hofman, A. *et al.* The Rotterdam Study: 2012 objectives and design update. *Eur. J. Epidemiol.* 26, 657–86 (2011).
17
Heckbert, S. R. *et al.* Antihypertensive treatment with ACE inhibitors or beta-blockers and risk of incident atrial fibrillation in a general hypertensive population. *Am. J. Hypertens.* 22, 538–44 (2009).
18
Smith, N. L. *et al.* Esterified estrogens and conjugated equine estrogens and the risk of venous thrombosis. *JAMA* 292, 1581–7 (2004).
19
Psaty, B. M. *et al.* The risk of myocardial infarction associated with antihypertensive drug therapies. *JAMA* 274, 620–5
20
Teumer, A. *et al.* Genome-wide association study identifies four genetic loci associated with thyroid volume and goiter risk. *Am. J. Hum. Genet.* 88, 664–73 (2011).
21
Inouye, M. *et al.* An immune response network associated with blood lipid levels. *PLoS Genet.* 6, e1001113 (2010).

To ensure reproducibility of the detected *trans*-eQTLs, we replicated *trans*-eQTLs from our discovery meta-analysis in 2 independent studies of peripheral blood gene expression: 52% in KORA F4 (n = 740 samples)[22] and 79% in BSGS (n = 862 samples)[23] (FDR < 0.05; Supplementary Figure 8). Irrespective of significance, 91% and 93% of all 1,513 significant *trans*-eQTL SNP-probe combinations showed consistent allelic direction in these replication cohorts compared with in the discovery analysis. A meta-analysis of the two replication studies improved replication rates: 89% of the 1,513 *trans*-eQTLs were significantly replicated (FDR <0.05), with 99.7% showing a consistent allelic direction. Irrespective of significance, 97% of the *trans*-eQTLs showed a consistent allelic direction in this replication meta-analysis (Supplementary Figure 8). We found that some *trans*-eQTLs could be detected in three cell type–specific data sets (283 monocyte samples[9], 282 B cell samples[9] and 608 HapMap lymphoblastoid cell line (LCL) samples[24]; Supplementary Figures 9 and 10). Despite the different tissues analyzed in these three studies, we were able to significantly replicate 7%, 4% and 2% of the *trans*-eQTLs (FDR <0.05), respectively. As 95% of the *trans*-eQTL SNPs explained less than 3% of the total expression variance (Supplementary Figure 11 and Supplementary Table 6), we lack statistical power to replicate most *trans*-eQTLs in these smaller replication cohorts.

We subsequently confined further analyses to 2,082 different SNPs that have been found to be associated with complex traits at genome-wide significance (trait-associated SNPs; reported $P < 5 \times 10^{-8}$; out of 4,542 unique SNPs that we tested). These 2,082 SNPs showed a significantly higher number of *trans*-eQTL effects compared with the 2,460 tested SNPs with reported disease associations at lower significance levels ($P = 8 \times 10^{-22}$; Supplementary Figure 12 and Supplementary Note): 254 of these 2,082 SNPs showed a *trans*-eQTL effect in the discovery analysis (reflecting 1,340 SNP-probe combinations; 1,201 of these were significantly replicated in blood, reflecting 233 different SNPs and 103 independent loci). For 671 of these 1,340 *trans*-eQTLs (50%), the trait-associated SNP (or a SNP in strong linkage disequilibrium, LD) was the strongest *trans*-eQTL SNP within the locus or was unlinked to the strongest *trans*-eQTL SNP (Supplementary Table 8 and Supplementary Note). The 2,082 trait-associated SNPs were 6 times more likely to cause *trans*-eQTL effects than were randomly selected SNPs (matched for distance to the gene and allele frequency; $P = 5.6 \times 10^{-49}$; Supplementary Figure 13 and Supplementary Note). SNPs associated with (auto)immune or hematological traits were twice as likely to underlie *trans*-eQTLs compared with other trait-associated SNPs ($P = 5 \times 10^{-25}$; Supplementary Note). Trait-associated SNPs that also caused *trans*-eQTLs affected the expression levels of nearby transcription factors in *cis* more frequently than trait-associated SNPs that did not affect genes in *trans* (Fisher's exact P = 0.032; Supplementary Note), suggesting that some *trans*-eQTLs arise owing to altered *cis* gene expression levels of nearby transcription factors.

We examined the genomic properties of the *trans*-eQTL SNPs (and their perfect proxies identified using data from the 1000 Genomes Project[25,26]): these SNPs were significantly enriched for mapping within microRNA (miRNA) binding sites (Fisher's exact

22
Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* 21, 48–54 (2013).
23
Powell, J. E. *et al.* The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS One* 7, e35430 (2012).
24
Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* 8, e1002639 (2012).
25
Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–73 (2010).
26
Patterson, K. 1000 genomes: a world of variation. *Circ. Res.* 108, 534–6 (2011).

Table 2. Complex traits where multiple unlinked SNPs affect the same downstream genes

Trait Type	Complex trait	At least two unlinked trait SNPs both affect gene(s):
Immune related traits	Type 1 diabetes	*GBP4, STAT1*
	Type 1 diabetes autoantibodies	*GBP4, STAT1*
	Celiac disease	*CXCR6, FYCO1*
	Multiple sclerosis	*CD5*
Blood pressure traits	Diastolic blood pressure	*LOC338758*
	Systolic blood pressure	*LOC338758*
Hematological traits	Hemoglobin	*ALAS2*
	Hematological Parameters	*FBXO7*
	F-cell distribution	*ESPN, PHOSPHO1, GNAS, TSPAN13, VWCE*
	Hematocrit	*ALAS2*
	Serum markers of iron status	*ALAS2*
	Red blood cell traits	*ALAS2*
	Serum iron levels	*ALAS2*
	Glycated hemoglobin levels	*ALAS2*
	Hematology traits	*ALAS2*
	Serum hepcidin	*ALAS2*
	Beta-thalassemia	*PHOSPHO1, VWCE, TSPAN13, ESPN*
	Hematological and biochemical traits	*AL109955.37-3, RBM38, TRIM58*
	Mean corpuscular hemoglobin	*ALAS2, C18orf10, DNAJB2, ESPN, HBM, KEL, PDZK1IP1, PIM1, PRDX5, RAP1GAP, UBXN6, VWCE, XK*
	Mean corpuscular volume	*ALAS2, B4GALT3, C18orf10, C1orf128, C22orf13, C5orf4, CCBP2, CSDA, DNAJB2, EIF2AK1, ESPN, FBXO7, HAGH, HBM, HPS1, KEL, KLC3, KRT1, LGALS3, MARCH8, MCOLN1, OSBP2, PDZK1IP1, PHOSPHO1, PIM1, PLEK2, PPP2R5B, PRDX5, PTMS, RAP1GAP, RIOK3, TGM2, TSTA3, UBXN6, VWCE, XK*
	Mean platelet volume	*ABCC3, AL353716.18, AQP10, C19orf33, C6orf152, CABP5, CTDSPL, CTTN, CXCL5, ESAM, F13A1, GNB5, GNG11, GP9, GUCY1A3, ITGA2B, ITGB5, LIMS1, LY6G6F, MMRN1, MPL, NRGN, PARVB, PRDX6, PTCRA, RAB27B, RBPMS2, SAMD14, SH3BGRL2, TSPAN9, VCL*

P < 0.05; Figure 1A). They mapped to regions in K562 (myeloid) and GM12878 (lymphoid) cell lines showing enrichment of histone enhancer signals (fold change > 2.5; Figure 1B) compared to the signals observed in six non-blood cell lines. Enhancer enrichment in myeloid and lymphoid cells supports the validity of our blood-derived *trans*-eQTLs. These results suggest that *trans*-eQTL effects are tissue specific, a notion that is supported by our inability to replicate a *trans*-eQTL that was previously identified in adipose tissue[13] for SNP rs4731702, associated with both type 2 diabetes (T2D) and lipid levels.

These *trans*-eQTLs can provide insight into the pathogenesis of disease. Although RNA microarray studies have identified dysregulated pathways for many complex diseases, it is often unclear whether associated SNPs first cause defects in the pathways whose dysregulation ultimately leads to disease or vice-versa. One example of this type of complex disease is SLE, which is an autoimmune disease causing inflammation and tissue damage. Individuals with SLE have increased type I interferon

27
Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl. Acad. Sci. U. S. A.* 100, 2610–5 (2003).
28
Bennett, L. *et al.* Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *J. Exp. Med.* 197, 711–23 (2003).

(IFN-α) levels, increased expression of IFN-α response genes[4,27,28] and decreased expression of the *C1Q* complement genes. We observed that four common SLE-associated variants affected IFN-α response genes in *cis* (*IRF5, IRF7, TAP2* and *PSMB9*; Supplementary Table 3). As most SLE-associated SNPs do not map near complement or IFN-α response genes, we assessed whether SLE-associated SNPs affect these genes in *trans*. This was the case for rs4917014, for which the SLE risk allele (rs4917014[T]; showing genome-wide significance in Asian populations and nominal significance in European populations[1,29]) not only increased expression of five different IFN-α response genes (*HERC5, IFI6, IFIT1, MX1* and *TNFRSF21*; Figure 2) but also decreased expression of three different probes in *CLEC10A*. We also observed a nominally significant association of rs4917014[T] with decreased expression of *C1QB* ($P = 5.2 \times 10^{-6}$; FDR = 0.28), encoding a subunit of the C1q complement complex, which has a protective role in lupus: complete deletion of the genes encoding the C1q subunits practically ensures the development of SLE[30,31]. *CLEC10A* and *CLEC4C* belong to the C-type lectin family, which includes mannose-binding lectins (MBLs).

Although, to our knowledge, *CLEC10A* and *CLEC4C* have not been studied in the context of SLE, the role of MBLs is similar to that of the C1q complex, and MBLs are a risk factor for the development of autoimmunity in humans and mice[3]. The rs4917014 *trans*-eQTLs replicated well in the peripheral blood and monocyte replication data sets and reinforce the role of altered expression of the IFN-α pathway, C-type lectin and C1Q genes in SLE. Individuals without SLE but who carry the rs497014[T] risk allele show these pathway alterations, indicating that these affected pathways are not solely a consequence of SLE but could precede SLE onset.

We investigated the underlying mechanisms of the effects exerted by rs4917014. *IKZF1* is the only gene overlapping the rs4917014 locus. As this gene encodes a transcription factor (Ikaros-family zinc finger 1), *cis*-regulatory effects of rs4917014 on *IKZF1* and consequent altered IKZF1 protein levels could constitute a mechanism for the detected *trans*-eQTL effects. However, because our meta-analysis did not initially detect a *cis*-eQTL on the Illumina probe for *IKZF1* located near the 5'-UTR of the gene, we investigated the 3'-UTR using Deep Serial Analysis of Gene Expression (DeepSAGE) next-generation RNA sequencing data from 94 peripheral blood samples[32]. The variant rs4917014[T] allele increased expression levels of the 3'-UTR of *IKZF1* (Spearman's correlation = 0.45; $P = 6.29 \times 10^{-6}$). Using Encyclopedia of DNA Elements (ENCODE) Project[33] chromatin immunoprecipitation and sequencing (ChIP-seq) data, we observed significantly increased IKZF1 protein binding within genomic locations corresponding with *trans*-eQTL–upregulated genes compared with all other genic DNA (Wilcoxon P value = 0.046) and with SLE *cis*-eQTL–upregulated genes outside of the *IKZF1* locus (Wilcoxon P value = 4.3×10^{-4}), thereby confirming the importance of *IKZF1* in SLE. *IKZF1* is also important for other phenotypes: rs12718597, an unlinked intronic variant within *IKZF1*, is associated with mean corpuscular volume (MCV)[34] and affects the expression of Illumina probe 4390086 near the 5'-end of *IKZF1* in *cis*. *Ikzf1* knockout mice show abnormal erythropoiesis[35], suggesting a causal role for human *IKZF1* in MCV

29 Wang, C. *et al.* Genes identified in Asian SLE GWASs are also associated with SLE in Caucasian populations. *Eur. J. Hum. Genet.* 21, 994–9 (2013).

30 McAdam, R. A., Goundis, D. & Reid, K. B. A homozygous point mutation results in a stop codon in the C1q B-chain of a C1q-deficient individual. *Immunogenetics* 27, 259–64 (1988).

31 Botto, M. *et al.* Homozygous C1q deficiency causes glomerulonephritis associated with multiple apoptotic bodies. *Nat. Genet.* 19, 56–9 (1998).

32 Zhernakova, D.V *et al.* DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* 9, e1003594 (2013).

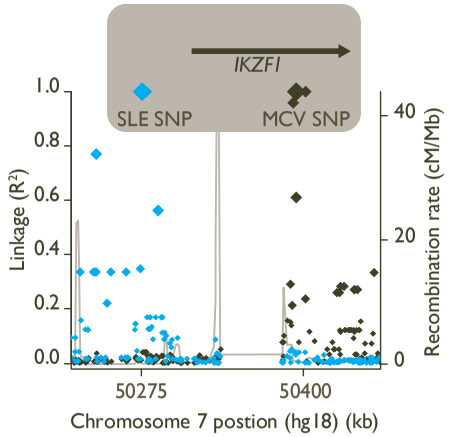
33 Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).

34 Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* 41, 1191–8 (2009).

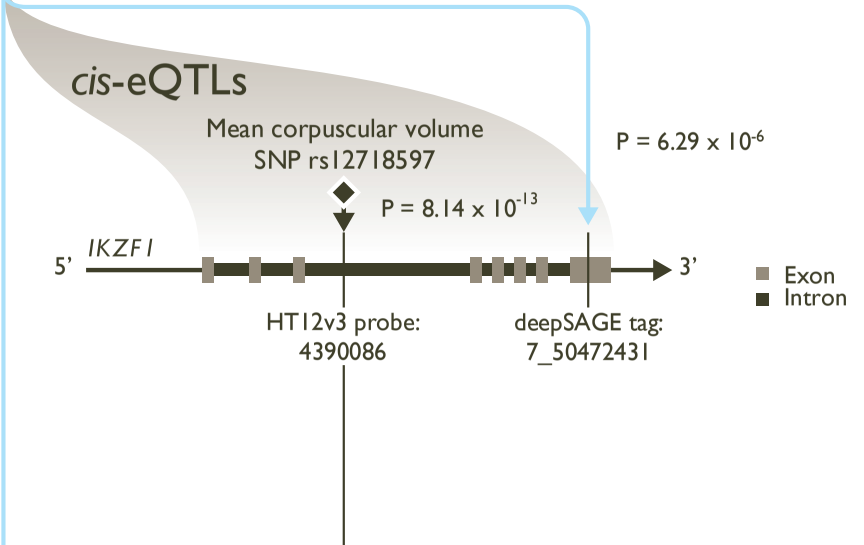
35 Wang, J. H. *et al.* Selective defects in the development of the fetal and adult lymphoid system in mice with an Ikaros null mutation. *Immunity* 5, 537–49 (1996).

Figure 2.

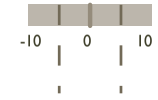
Independent *trans*-eQTL effects emanating from the *IKZF1* locus. SNP rs4917014, associated with SLE, and unlinked SNP rs4917014, associated with MCV, both affect the expression of *IKZF1* in *cis*. rs12718597 affects 50 genes in *trans* (mostly involved in hemoglobin metabolism), and rs4917014 affects 8 different genes in *trans*: the rs4917014[T] risk allele is associated with increased expression of genes involved in the type I interferon response.
At a somewhat lower significance threshold (FDR = 0.28), rs4917014[T] is associated with decreased complement *C1QB* expression. Both processes are hallmark features of SLE.



Systemic lupus erythematosus allele **rs4917014 (T)** affects ***IKZF1*** in *cis* and many genes in *trans*:



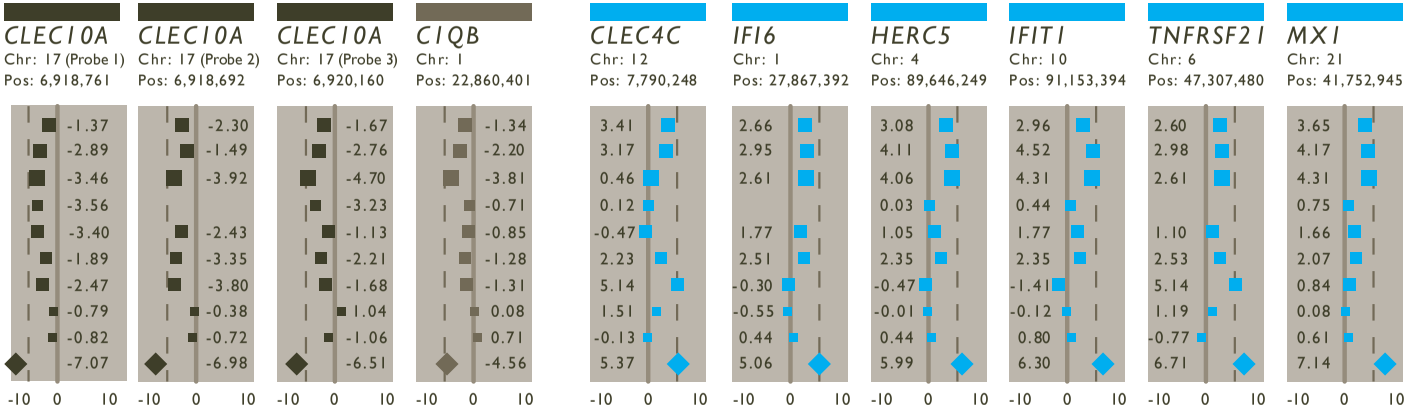
■ Increases expression (FDR < 0.05)
■ Decreases expression (FDR < 0.05)
* Significant replication (FDR < 0.05)
NS Non-significant replication (FDR > 0.05)



MCV *trans*-eQTL effects

Genes involved in hemoglobin and erythrocyte metabolic processes:
AC010679.1, ALDH5A1, AP2S1, B4GALT3, C19orf62, C1orf128, C22orf13, C5orf4, CCBP2, CSDA, E2F2, EIF2AK1, EIF3S9, FAM104A, FBXO7, GCAT, GPR146, HAGH, HEMGN, HK1, HPS1, KCNH2, KLC3, KRT1, LGALS3, MAP2K3, MARCH8, MCOLN1, MSI2, OSBP2, PDLIM7, PFDN5, PLEK2, PPP2R5B, PTMS, RAP1GAP, RIOK3, RP11-529I10.4, RPIA, SESN3, SIAH2, SLC38A5, SLC6A8, SLC7A5, STOML2, TFDP1, TGM2, TMEM86B, TSTA3, VWCE

SLE *trans*-eQTL effects



Discovery cohorts:																				
EGCUT		-1.37		-2.30		-1.67		-1.34		3.41		2.66		3.08		2.96		2.60		3.65
SHIP-TREND		-2.89		-1.49		-2.76		-2.20		3.17		2.95		4.11		4.52		2.98		4.17
Fehrmann HT12		-3.46		-3.92		-4.70		-3.81		0.46		2.61		4.06		4.31		2.61		4.31
Fehrmann H8v2		-3.56				-3.23		-0.71		0.12				0.03		0.44				0.75
Rotterdam Study		-3.40		-2.43		-1.13		-0.85		-0.47		1.77		1.05		1.77		1.10		1.66
DILGOM		-1.89		-3.35		-2.21		-1.28		2.23		2.51		2.35		2.35		2.53		2.07
InCHIANTI		-2.47		-3.80		-1.68		-1.31		5.14		-0.30		-0.47		-1.41		5.14		0.84
HVH HT12v3		-0.79		-0.38		1.04		0.08		1.51		-0.55		-0.01		-0.12		1.19		0.08
HVH HT12v4		-0.82		-0.72		-1.06		0.71		-0.13		0.44		0.44		0.80		-0.77		0.61
Meta-analysis		-7.07		-6.98		-6.51		-4.56		5.37		5.06		5.99		6.30		6.71		7.14
Replication cohorts:																				
B-Cells	NS	-0.25	NS	-2.59	NS	-0.32	NS	0.38	NS	1.42	NS	2.78	NS	2.93	NS	2.42	NS	2.18	*	3.41
Monocytes	*	-3.45	*	-3.45	NS	-2.70	NS	1.78	NS	0.55	*	4.13	*	4.13	*	4.03	NS	1.90	*	3.83
Peripheral blood (KORA F4)	NS	-2.15	*	-2.83	NS	-1.47	NS	-0.22	NS	0.57	*	3.57	*	3.60	*	3.46	*	4.37	*	3.96
Peripheral blood (BSGS)	NS	1.89	NS	-0.38	*	-2.21	*	-3.41	NS	0.12	*	3.48	*	4.10	*	4.44	*	4.31	*	4.13

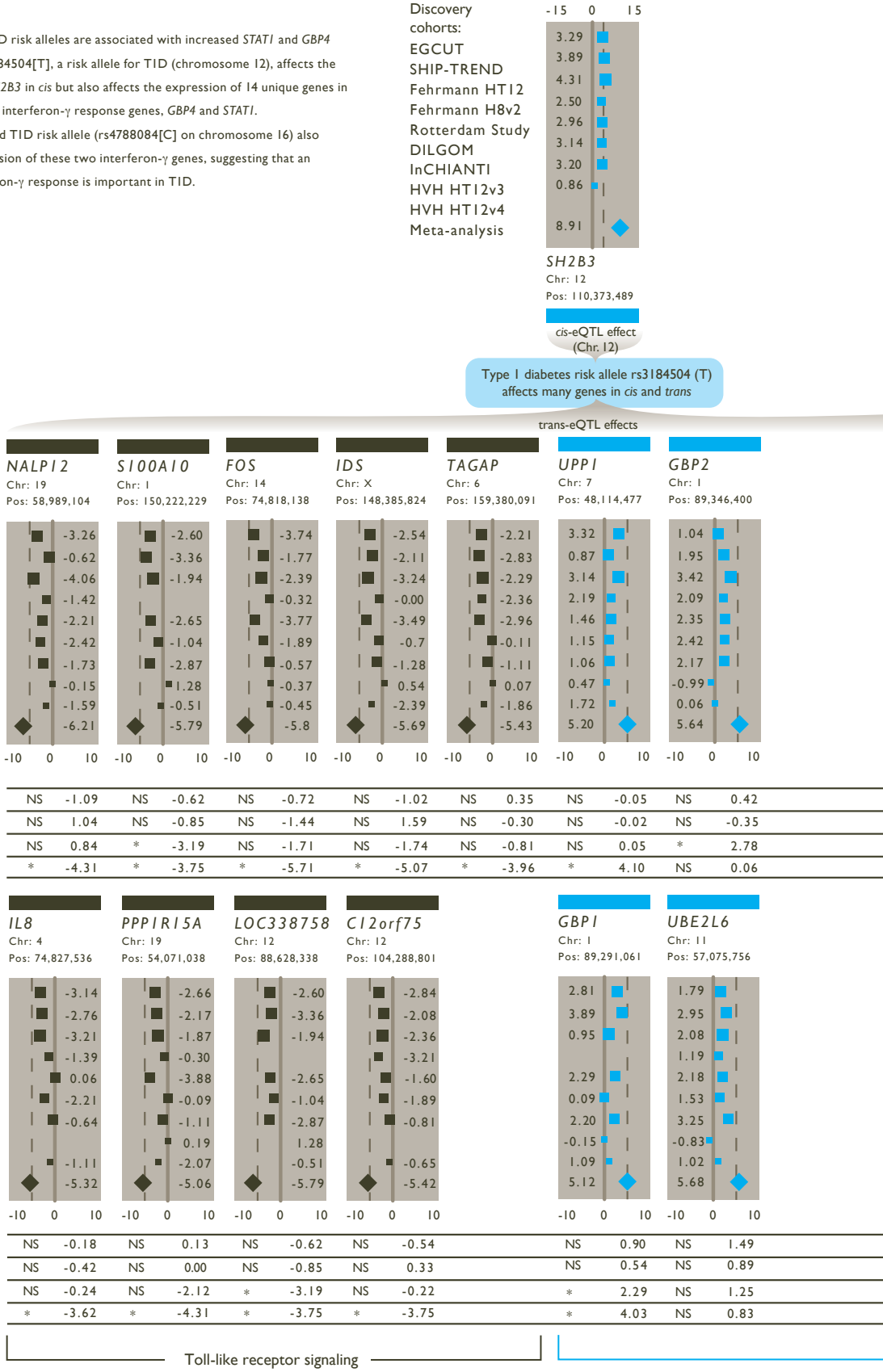
Genes involved in complement

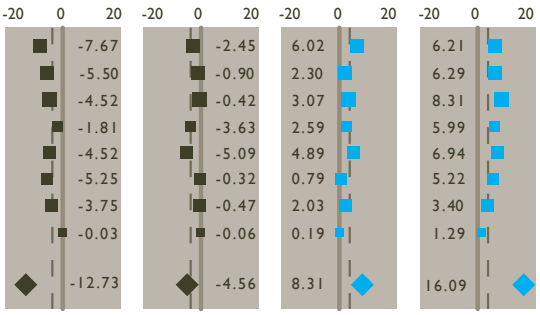
Type I Interferon response genes
Enrichment of IKZF1 binding (Wilcoxon P = 0.05)

Figure 3.

Two unlinked T1D risk alleles are associated with increased *STAT1* and *GBP4* expression. rs3184504[T], a risk allele for T1D (chromosome 12), affects the expression of *SH2B3* in *cis* but also affects the expression of 14 unique genes in *trans*, including 2 interferon-γ response genes, *GBP4* and *STAT1*.

Another, unlinked T1D risk allele (rs4788084[C] on chromosome 16) also increases expression of these two interferon-γ genes, suggesting that an elevated interferon-γ response is important in T1D.



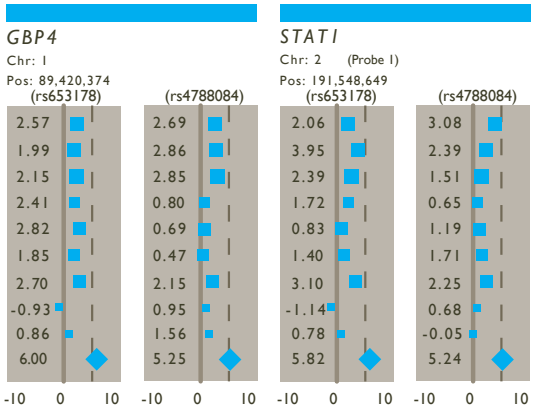


Discovery cohorts:
EGCUT
SHIP-TREND
Fehrmann HT12
Fehrmann H8v2
Rotterdam Study
DILGOM
InCHIANTI
HVH HT12v3
HVH HT12v4
Meta-analysis

cis-eQTL effects (Chr. 16)

Type 1 diabetes risk allele rs4788084 (C) affects ***GBP4*** and ***STAT1*** in *trans*

trans-eQTL effects



Increases expression (FDR < 0.05)
Decreases expression (FDR < 0.05)
* Significant replication (FDR < 0.05)
NS Non-significant replication (FDR > 0.05)

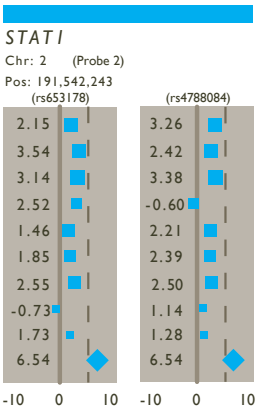
Direction of effect (Z-score)
FDR 0.05 threshold

Discovery cohorts:
EGCUT
SHIP-TREND
Fehrmann HT12
Fehrmann H8v2
Rotterdam Study
DILGOM
InCHIANTI
HVH HT12v3
HVH HT12v4
Meta-analysis

Replication cohorts:

NS	-0.87	NS	-1.05	NS	0.65	NS	1.96
NS	-0.17	NS	0.13	NS	0.39	NS	-0.23
*	3.68	*	2.23	*	2.59	NS	1.41
*	4.69	*	4.69	*	3.99	*	4.74

Replication cohorts:
B-Cells
Monocytes
Peripheral blood (KORA)
Peripheral blood (BSGS)



Discovery cohorts:
EGCUT
SHIP-TREND
Fehrmann HT12
Fehrmann H8v2
Rotterdam Study
DILGOM
InCHIANTI
HVH HT12v3
HVH HT12v4
Meta-analysis

Replication cohorts:

NS	0.72	NS	1.42
NS	-0.47	NS	0.96
*	3.60	*	2.83
*	3.72	*	4.69

Replication cohorts:
B-Cells
Monocytes
Peripheral blood (KORA)
Peripheral blood (BSGS)

Response to interferon-gamma

as well. However, although rs12718597[A] was associated in *trans* with the upregulation of 31 genes and with the downregulation of 19 genes, none of the SLE *trans*-regulated genes overlapped with the MCV *trans*-regulated genes. The latter were mainly involved in hemoglobin metabolism and did not show increased IKZF1 binding (Wilcoxon P value = 0.35). In summary, these results indicate that *IKZF1* has multiple functions and that different SNPs near *IKZF1* elicit function-specific effects.

We identified other *trans*-eQTLs showing similar phenomena. For example, rs174546 (located in the 3'-UTR of *FADS1* and associated with metabolic syndrome[36] and with low-density lipoprotein (LDL) and total cholesterol levels[37,38]) affected the expression of *TMEM258, FADS1* and *FADS2* in *cis* and the expression of *LDLR* in *trans* (Supplementary Figure 14). *LDLR* encodes the LDL receptor and contains common variants that are also associated with lipid levels[38]. *LDLR* gene expression levels correlated negatively ($P < 3.0 \times 10^{-4}$) with total, high-density lipoprotein (HDL) and LDL cholesterol levels in the tested cohorts (Rotterdam Study and EGCUT; Supplementary Table 9), indicating that peripheral blood is a useful tissue for gaining insight into the downstream effects of lipid-regulating SNPs.

For 21 different complex traits, at least 2 unlinked variants that have been associated with these diseases affected exactly the same gene in *trans* (compared with 1 complex trait similarly affected by variants from equally sized but permuted lists of *trans*-eQTLs; Table 2, Supplementary Figure 15 and Supplementary Table 10). Although most of these traits are hematological (for example, mean platelet volume or serum iron levels), we also observed this convergence for blood pressure, celiac disease, multiple sclerosis and type 1 diabetes (T1D).

rs3184504 (located in an exon of *SH2B3*) and its proxy rs653178 (located in an intronic region of *ATXN2* on chromosome 12) have been associated with several autoimmune diseases, including T1D[39,40] and the production of autoantibodies therein[39,40], celiac disease[8,41], hyperthyroidism[42], vitiligo[43] and rheumatoid arthritis[41], as well as with other complex traits such as blood pressure[44,45], chronic kidney disease[46] and eosinophil counts[47]. We observed a *cis*-eQTL effect for this SNP on *SH2B3* (FDR < 0.05) and *trans*-eQTL effects on 14 genes (FDR < 0.05; Figure 3), all of which are highly expressed in neutrophils. Because the *trans*-eQTLs effects could be explained by known effect of rs3184504 on differences in cell count proportions[47], we correlated the expression levels of *trans*-regulated genes with cell counts in two cohorts (Rotterdam Study and EGCUT) but did not observe significant correlations (Supplementary Table 9). The identified *trans*-eQTLs describe different biological functions: the T1D risk allele rs3184504[T] was associated with decreased expression of nine genes, most of which are involved in Toll-like receptor signaling[48] (*C12orf75, FOS, IDS, IL8, LOC338758, NALP12, PPP1R15A, S100A10* and *TAGAP*) and with increased expression of five genes involved in the interferon-γ response (*GBP2, GBP4, STAT1, UBE2L6* and *UPP1*). We observed that another T1D risk allele, rs4788084[C][39,40] on chromosome 16, was also associated with increased expression of *GBP4* and *STAT1*, showing how different T1D risk alleles converge, with both alleles causing an increase in the expression of interferon-γ response genes.

36 Zabaneh, D. & Balding, D. J. A genome-wide association study of the metabolic syndrome in Indian Asian men. *PLoS One* 5, e11961 (2010).

37 Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 41, 35–46 (2009).

38 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–13 (2010).

39 Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* 41, 703–7 (2009).

40 Plagnol, V. *et al.* Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet.* 7, e1002216 (2011).

41 Zhernakova, A. *et al.* Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci. *PLoS Genet.* 7, 13 (2011).

42 Eriksson, N. *et al.* Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS One* 7, e34442 (2012).

43 Jin, Y. *et al.* Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nat. Genet.* 44, 676–80 (2012).

44 Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* 41, 666–76 (2009).

45 Wain, L.V *et al.* Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat. Genet.* 43, 1005–11 (2011).

46 Köttgen, A. *et al.* New loci associated with kidney function and chronic kidney disease. *Nat. Genet.* 42, 376–84 (2010).

47 Gudbjartsson, D. F. *et al.* Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.* 41, 342–7 (2009).

48 Rotival, M. *et al.* Integrating genome-wide genetic variations and monocyte expression data reveals *trans*-regulated gene modules in humans. *PLoS Genet.* 7, e1002367 (2011).

In summary, our eQTL meta-analysis identified and replicated downstream effects for 233 trait-associated SNPs. Our analyses show that *trans*-eQTL mapping in blood for lipid-regulatory and immune-mediated disease variants yields insights into downstream pathways that are biologically meaningful. Future, larger-scale *trans*-eQTL analyses in blood will likely uncover many more of these regulatory relationships.

Methods

Study populations

We performed a whole-genome eQTL meta-analysis of 5,311 samples from peripheral blood divided over a total of 9 data sets from 7 cohorts, including EGCUT[14] (n = 891), InCHIANTI[15] (n = 611), Rotterdam Study[16] (n = 762), Fehrmann[5] (n = 1,240 on the Illumina HT12v3 platform and 229 on the Illumina H8v2 platform), HVH[17–19] (n = 43 on the Illumina HT12v3 platform and 63 on the Illumina HT12v4 platform), SHIP-TREND[20] (n = 963) and DILGOM[21] (n = 509). Gene expression data for each data set were obtained by isolating RNA using either PAXGene (Becton Dickinson) or Tempus (Life Technologies) tubes and then hybridizing RNA to Illumina whole-genome Expression BeadChips (HT12v3, HT12v4 or H8v2 arrays). Gene expression platforms were harmonized by matching probe sequences across the different platforms. Mappings for these sequences were obtained by mapping the sequences against Build 36 of the human genome (Ensembl Build 54, hg18) using the BLAT, BWA and SOAPv2 sequence alignment programs. Highly stringent alignment criteria were used to ensure that probes mapped unequivocally to a single genomic position. Genotype data were acquired using different genotyping platforms and were harmonized by imputation, using the HapMap 2 CEU population as a reference[49]. Each data set was individually checked for sample mix-ups using *MixupMapper*[50]. For a full description of the individual data sets, the results of the sample mix-up analysis, specifics on the gene expression platforms used and probe mapping and filtering procedures, see the Supplementary Note.

Gene expression normalization

Gene expression data were quantile normalized to the median distribution and were subsequently $\log_2$ transformed. Probe and sample means were centered to zero. Gene expression data were then corrected for possible population structure through the removal of four multidimensional scaling components using linear regression. We reasoned earlier that normalized gene expression data still contain large amounts of non-genetic variation[5]. Therefore, after correction for population stratification, we performed principal-component analysis (PCA) on the sample correlation matrix. We performed a separate QTL analysis for each principal component to ascertain whether genetic variants could be detected that affected each principal component. If we found an effect on the principal component, we did not correct the expression data for this component to ensure that we would not unintentionally remove genetic effects from the expression data. We established the significance of these associations by controlling the FDR, testing each association against a null distribution created by repeating the analysis 100

49 The International HapMap Project. *Nature* 426, 789–796 (2003).

50 Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* 27, 2104–2111 (2011).

times (permuting the sample labels for each iteration[51]). Principal components that did not show significance at the FDR threshold of 0.0 were removed from the gene expression data by linear regression. In all but 2 very small data sets, the first 40 principal components were removed (excluding those components for each cohort that showed a QTL effect). We observed that the removal of these 40 components resulted in the identification of the highest number of eQTLs in each data set. Although principal-component correction might remove some eQTL effects, we observed that the majority of *trans*-eQTL effects (95% when removing 35 principal components and 90% when removing 45 principal components) were independent of the number of principal components removed (Supplementary Figure 16).

eQTL mapping

After normalization of the data, we performed both *cis*- and *trans*-eQTL mapping. eQTLs were deemed *cis*-eQTLs when the distance between the SNP chromosomal position and the probe midpoint was less than 250 kb, whereas eQTLs with a distance greater than 5 Mb were defined as *trans*-eQTLs. Only SNPs with a minor allele frequency (MAF) of > 0.05 and a Hardy-Weinberg equilibrium P value of > 0.001 were included in the analyses. Because most cohorts had generated gene expression data using the HT12v3 platform, we chose to only include probes that were present on this platform. We only tested SNP-probe pairs when the SNP passed quality control in at least three cohorts. Furthermore, to address issues with respect to computational time and multiple testing, we confined our *trans*-eQTL analysis to those SNPs present in the Catalog of Published GWAS (see URLs; accessed 16 July 2011). We reasoned that, for genes with strong *cis*-eQTL effects, a *cis*-eQTL effect might obscure the detectability of *trans*-eQTLs. Therefore, we used linear regression to remove *cis*-eQTL effects before *trans*-eQTL mapping and observed a 12% increase in the number of detected *trans*-eQTLs (Supplementary Figure 17). For each cohort, eQTLs were mapped using a Spearman's rank correlation on imputed genotype dosages. We used a weighted z-score method for subsequent meta-analysis[52]. To generate a realistic null distribution, we permuted the sample identifiers of the expression data and repeated this analysis ten times (Supplementary Figure 18). In each permutation, the sample labels were permuted. We then corrected for multiple testing by setting the FDR at 0.05, testing each P value in the real data against a null distribution created from the permuted data sets[51] (Supplementary Note). It has been suggested that false-positive eQTL effects can arise owing to polymorphisms in the probe sequences[53,54]. Therefore, we tested whether a significant *cis*-eQTL SNP was in LD ($r^2 > 0.2$) with any SNP in the *cis* probe sequence, using the Western European subpopulations of the 1000 Genomes Project[25] (2011-05-21 release; 286 individuals, excluding Finnish individuals) as a reference. If we observed this to be the case, the respective *cis*-eQTLs were removed. Furthermore,for each *trans*-eQTL, we investigated whether portions of the probe sequence could be mapped to the vicinity of the *trans*-eQTL SNP (which would imply a *cis*-eQTL rather than a *trans*-eQTL effect). For this analysis, we tried to map the *trans*-eQTL probe sequences, using very permissive settings, within a 5-Mb window centered on the *trans*-eQTL SNP. SNP-probe combinations where at least 15 bp of the probe mapped within this 5-Mb window were deemed false positives and

51 Breitling, R. *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 4, e1000232 (2008).

52 Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* 18, 1368–73 (2005).

53 Alberts, R. *et al.* Sequence polymorphisms cause many false cis-eQTLs. *PLoS One* 2, e622 (2007).

54 Benovoy, D., Kwan, T. & Majewski, J. Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. *Nucleic Acids Res.* 36, 4417–23 (2008).

were removed from further analysis. After this filtering, we recalculated the FDR for both the *cis*- and *trans*-eQTL results.

Trans-eQTL replication

Replication of the *trans*-eQTL results was carried out in 5 independent data sets from 4 cohorts, including data obtained from LCLs (HapMap 3, n = 608)[24], B cells and monocytes (Oxford, n = 282 and 283, respectively)[9] and whole peripheral blood (KORA F4, n = 740 and BSGS, n = 862)[22,23]. All cohorts applied the same methodology as used in the discovery phase to normalize gene expression data, check for sample mix-ups and perform *trans*-eQTL mapping, including ten permutations to establish the FDR threshold at 0.05. Finally, we performed a sample size–weighted z-score meta-analysis on the two peripheral blood replication cohorts (KORA F4 and BSGS). Further details on these data sets can be found in the Supplementary Note.

Enhancer enrichment and functional annotation

To determine whether the significant *trans*-eQTL SNPs were enriched for functional regions on the genome, we annotated the *trans*-eQTL SNPs using SNPInfo[55], SNPNexus[56,57] and HaploReg[58], which integrate multiple data sources (such as the ENCODE Project[33], Ensembl[59] and several miRNA databases). We limited these analyses to those *trans*-eQTL SNPs that were previously shown to be associated with complex traits at genome-wide significance (trait-associated SNPs; reported $P < 5 \times 10^{-8}$). These SNPs were subsequently pruned (using the --clump command in PLINK with $r^2 < 0.2$). We used permuted *trans*-eQTL data to generate realistic null distributions for each of these tools: we selected equally sized sets of unlinked SNPs ($r^2 < 0.2$ in the Western European subpopulations of the 1000 Genomes Project[25], 2011-05-21 release; 286 individuals, excluding Finnish individuals) that showed the highest significance in the permuted data, ensuring that only trait-associated SNPs were included in the null distribution, as it is known that trait-associated SNPs in general already have different functional properties than randomly selected SNPs[60] (for example, trait-associated SNPs typically map in closer proximity to genes than randomly selected SNPs). We also ensured that none of the SNPs in the null distribution were affecting genes in *trans* or were linked to those SNPs ($r^2 < 0.2$ in 1000 Genomes Project data). We then identified perfect proxies ($r^2 = 1.0$ in 1000 Genomes Project data). For SNPInfo and SNPNexus, we calculated the enrichment for each functional category using a Fisher's exact test. We examined enhancer enrichment in nine different cell types using HaploReg, averaging enhancer enrichment over the ten permutations.

Convergence analysis

We determined which unlinked trait-associated SNPs showed eQTL effects on exactly the same gene: for each trait, we analyzed the SNPs that are known to be associated with the trait and assessed whether any unlinked SNP pair ($r^2 < 0.2$; distance between SNPs of >5 Mb) showed a *cis*- and/or *trans*-eQTL effect on exactly the same gene, as previously described[5]. To determine whether the number of traits for which we observed this phenomenon was higher than expected by chance, we repeated this analysis 20 times, each time using a different set of permuted *trans*-eQTLs, equal in size to the non-permuted set of *trans*-eQTLs.

55 Xu, Z. & Taylor, J. A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* 37, W600–5 (2009).

56 Chelala, C., Khan, A. & Lemoine, N. R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* 25, 655–61 (2009).

57 Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res.* 40, W65–70 (2012).

58 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–4 (2012).

59 Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* 40, D84–90 (2012).

60 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9362–7 (2009).

SLE IKZF1 ENCODE ChIP-seq analysis

We used IKZF1 ChIP-seq signal data obtained from the ENCODE Project[33] (IKZF1 ChIP-seq data acquired and processed by UCSC, ENCODE; March 2012 Freeze). For every human gene, we determined the average signal (corrected for gene size and bias in GC content) and performed a Wilcoxon Mann-Whitney test to determine whether the upregulated genes (*MX1, TNFRSF21, IFIT1-LIPA, HERC5, CLEC4C* and *IFI6*) showed a higher ChIP-seq signal than the average signal for all other human genes.

URLs

Catalog of Published GWAS (16 July 2011), http://www. genome.gov/gwastudies/.

Accession codes

We have made a browser available for all significant *trans*-eQTLs and *cis*-eQTLs at http://www.genenetwork.nl/bloodeqtlbrowser. This browser also provides all *trans*-eQTLs that we detected at a somewhat less stringent FDR of 0.5 to enable more in-depth post hoc analyses. Gene expression data are available for download at the Gene Expression Omnibus (GEO) (GSE36382, GSE20142, GSE20332, GSE33828, GSE33321, GSE47729; GSE48348 and GSE48152) and ArrayExpress (E-TABM-1036, E-MTAB-945 and E-MTAB-1708).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

Acknowledgments

Acknowledgments for each participating cohort can be found in the Supplementary Note.

Author Contributions

Experiment design and method development: H.-J.W., M.J.P., T.E., H.Y., C.S., J. Kettunen, M.W.C., B.P.F., K. Schramm, J. Karjalainen, T.L., Y.L., R.C.J., B.M.P., S.R., A.T., T.M.F., A.M., J.B.J.M. and L. Franke.

Reviewing and editing of the manuscript: H.-J.W., M.J.P., T.E., H.Y., C.S., J. Kettunen, M.W.C., B.P.F., A.Z., A.G.U., A.H., F.R., V.S., J.B., T.L., Y.L., R.C.J., P.M.V., J.C.K., B.M.P., S.R., A.T., T.M.F., A.M., J.B.J.M. and L. Franke.

Data collection: D.V.Z., J.H.V., J. Karjalainen, S.W., F.R., P.A.C.t.H., E.R., K.F., M. Nelis, L.M., D.M., L. Ferrucci, A.B.S., D.G.H., M.A.N., G.H., M. Nauck, D.R., U.V., M.P., A.S.-D., S.A.G., D.A.E., G.W.M., S.M., H.P., C.H., M.R., H.G., T.M., K. Strauch and L.H.V.d.B.

Replication of *trans*-eQTL results: B.P.F., K. Schramm, J.E.P., P.M.V. and J.C.K.

Competing financial interests

The authors declare no competing financial interests.
Reprints and permissions information is available online at
http://www.nature.com/reprints/index.html.

Part

3

Discussion

Part 1 – A broader perspective on the work described in this thesis

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex traits and diseases[1]. However, the genome-wide significant variants identified per trait so far explain only a limited fraction of the heritability. One explanation has been that the common variants identified might tag low frequency variants with a large effect size. However, this has only been shown for a few loci. This suggests that most of the associated variants are tagging common, low risk, causal variants[2].

The observation that the great majority of these variants are not causing changes to the protein structure (i.e. are non-coding) suggests the majority are regulatory variants. To identify such regulatory effects, it is very valuable to study how these genetic variants affect molecular phenotypes, by conducting quantitative trait locus (QTL) mapping. In the various steps from DNA to eventual disease phenotype, transcription is the first (intermediate) molecular phenotype. Since gene expression levels can now be easily measured using microarray-based technologies or through RNA-sequencing, expression QTL (eQTL) mapping is now often used to better interpret the effects of disease-associated genetic variants. Such eQTLs studies are usually conducted in *cis* (see Box 1), where only single nucleotide polymorphisms (SNPs) and genes are tested that map in close proximity to the same chromosome. *Cis*-eQTLs are very useful for pinpointing potential causal genes within disease loci and have been found in large quantities by several studies[3–24]. Distant effects, involving SNPs and genes that map at long distances from each other (i.e. *trans*-eQTLs; see Box 1) have been reported by only a few studies[9–12,18,25–28]. However, they are especially interesting because the discovery that a SNP is affecting a gene on a different chromosomes implies a biological relationship. *Trans*-eQTL mapping thus enables the detection of previously unknown downstream effects of disease-associated and other SNPs. Such knowledge can provide important insight into pathways that are causally involved in disease pathogenesis. However, detecting these *trans*-eQTLs has been difficult in the past, because they can be strongly cell-type-specific[9]. Because these *trans*-eQTLs have typically been studied in whole tissues (i.e. mixtures of different cell types), the observed effect sizes are typically small, and because many statistical tests need to be conducted, it has been difficult to identify them.

The work described in this thesis had two main aims: 1) to describe computational methods that can be used to improve the statistical power to detect *cis*- and *trans*-eQTLs, and 2) to identify previously unknown *cis*- and *trans*-eQTLs, in order to identify the downstream effects for many disease-associated genetic variants.

We have described several computational methods that improve power to detect eQTLs by: 1) correcting the eQTL datasets for the presence of accidental sample mix-ups,
2) correcting the gene expression for non-genetic effects,
3) increasing the sample size through meta-analysis, and
4) by applying an interaction model to identify cell-type-specific effects.

1
Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9362–7 (2009).
2
Hunt, K. A. *et al.* Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* 498, 232–5 (2013).
3
Bullaughey, K., Chavarria, C. I., Coop, G. & Gilad, Y. Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Hum. Mol. Genet.* 18, 4296–303 (2009).
4
Cvejic, A. *et al.* SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* 45, 542–5 (2013).
5
Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246–50 (2009).
6
Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* 39, 1202–7 (2007).
7
Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302 (2010).
8
Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* 452, 423–8 (2008).
9
Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 44, 502–10 (2012).
10
Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* 44, 1084–9 (2012).
11
Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* 8, e1003029 (2012).
12
Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* 7, e1002078 (2011).
13
Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–11 (2013).
14
Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 152, 633–41 (2013).
15
Liang, L. *et al.* A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.* 23, 716–26 (2013).
16
Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* 39, 1494–9 (2007).

Box 1. The confusing *cis* and *trans* terminology

Traditionally, a *cis*-regulatory element has been assumed to indicate that a certain region on a DNA molecule is regulating the expression of a gene that is located on that same DNA molecule. A *trans*-regulatory element can regulate the expression of a gene that is located on a different DNA molecule.

eQTL studies also use the distinction between *cis* and *trans*, although with a completely different meaning: since eQTL studies have previously been conducted using microarray-based gene expression measurements, it was not possible to determine whether a certain allele (located on a certain DNA molecule) was affecting the expression of a gene that is located on that same DNA molecule (i.e. haplotype).

The distinction between *cis* and *trans*-eQTLs is solely based on the distance between the genetic variant and the gene: if the genetic variant maps within 1 megabase of the gene, it is typically considered to be a *cis*-eQTL effect. In this thesis, we have shown that 95% of the *cis*-eQTL effects can be found within 50 kb of the genotype (Chapter 7). *Trans*-eQTL effects are considered to reflect eQTLs that involve a genetic variant and gene that map to different chromosomes or that map at a distance of at least five megabases, to exclude any possibility of long-range linkage disequilibrium.

A *cis*-eQTL is thus, strictly speaking, not necessarily a *cis*-regulatory effect (although a *cis*-regulatory effect is a *cis*-eQTL). Luckily, new technologies can resolve this confusion: RNA-sequencing permits the detection of allele-specific expression effects, and phasing algorithms now exist that can reconstruct large haplotypes (with only a few switching errors). Combined, this now makes it possible to determine whether a certain allele has a truly *cis*-regulatory effect on the gene that resides on the same DNA molecule.

To demonstrate the validity of these methods, we combined them and conducted several eQTL mapping studies in order to detect *cis*- and *trans*-eQTLs for: 1) SNPs influencing non-coding RNA gene expression levels, 2) SNPs influencing alternative poly-adenylation using RNA-seq, and 3) SNPs influencing downstream gene expression levels (*trans*-eQTLs).

In the first part of this thesis we provided a number of tools that can aid in increasing statistical power: in Chapter 2, we described the *MixupMapper* algorithm, which can identify the presence of sample mix-ups in eQTL datasets and can also correct these errors. We applied this method to several publicly available eQTL datasets and found that approximately 3% of all samples have been accidentally mixed-up. By correcting these sample mix-ups, we observed a 15% increase in the number of detectable eQTLs in these public datasets, and we could explain up to 1.24-fold more of the heritability for simulated complex traits. Since then, it has been shown that it is also possible to predict genotypes from gene expression or other high-throughput datasets (e.g. proteomic or metabolomics datasets)[29]. Consequently, this implies that by using many different types of quantitative data that have been deposited in public databases, it may be possible to identify individuals to a considerable extent, which may result in privacy issues[30]. These issues are even more relevant for RNA-seq based expression studies, since genotypes can be derived directly from raw RNA-sequence reads[31]. Researchers who aim to deposit raw RNA-sequence reads of human samples in public repositories, such as the European Nucleotide Archive, need to ensure they have the correct clearance and particularly that they have fully informed consent from their patients.

17 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–7 (2010).

18 Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* 21, 48–54 (2013).

19 Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* 7, e1002003 (2011).

20 Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* 39, 1217–24 (2007).

21 Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* 8, e1002639 (2012).

22 Zhong, H. *et al.* Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* 6, e1000932 (2010).

23 Zhang, W. *et al.* Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum. Genet.* 125, 81–93 (2009).

24 Zeller, T. *et al.* Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One* 5, e10693 (2010).

In Chapter 3, we presented methods to remove non-genetic variation in gene expression, by applying principal component analysis (PCA). We showed that removing principal components (PCs) can yield a more than two-fold increase in the number of detectable eQTLs. However, correcting for principal components can also remove genetic variation, which is especially relevant for *trans*-eQTL SNPs that affect multiple genes. Our method therefore tests each PC for genetic association and does not remove those components that show a genome-wide significant genetic association. However, the PCs that are used as covariates might still capture small genetic effects that are not genome-wide significant. Recently, other normalization strategies have therefore been suggested, including supervised normalization and correction for measured covariates (such as age, gender, cell counts)[32]. A comparison between such normalization methods showed that covariate correction outperforms the other normalization strategies tested, although PC correction appeared to be an alternative that performed well if such covariates were unavailable or unknown[33]. A recently published method, called HCP (hidden covariates with prior) combines both methods by using Bayesian statistics to define the hidden covariates, which can then be corrected through latent factor analysis[34].

Another method to improve the power to detect genetic effects is to increase the sample size. In Chapter 3, we described methodology to effectively meta-analyze eQTL datasets and showed that a *cis*- and *trans*-eQTL meta-analysis provides more statistical power with biologically meaningful results. We therefore used this meta-analysis method, the PC correction method and the sample mix-up detection method in the other work described in this thesis.

In Chapter 4 we presented a method that is able to predict the cell-type-specificity of eQTLs from gene expression data that was obtained from tissues composed of mixtures of cell types (in our case, whole blood). As samples are more easily collected from whole tissues than isolated cell types, our method enables the collection of the large number of samples that is required to detect small effects (such as cell-type-specific *trans*-eQTL effects). The interaction effect between genotype, cell count estimate and gene expression that was captured by our method was generally smaller than the main eQTL effects, requiring us to combine multiple datasets through meta-analysis. We were able to validate our predictions in eQTL datasets of purified blood cell types, thereby indicating the usefulness of our method. Our method is especially helpful since eQTL datasets on individual cell types are not yet available for many different cell types. Although we have provided a proof-of-principle in whole blood, our method is also applicable to other compound tissues: if, for the individual cell types that comprise this tissue, we know genes that are markers for these cell types, it possible to estimate the abundances for each cell type. With more public cell-type- and tissue-specific eQTL datasets now becoming available (through projects such as GTeX and IMMVAR), it will become possible to apply our method to many more tissues in the near future.

In part 2 of this thesis, we have shown that the computational methods from part 1 can be effectively used to identify eQTLs

25
Rotival, M. *et al.* Integrating genome-wide genetic variations and monocyte expression data reveals *trans*-regulated gene modules in humans. *PLoS Genet.* 7, e1002367 (2011).

26
Small, K. S. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat. Genet.* 43, 561–4 (2011).

27
Petretto, E. *et al.* Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet.* 2, e172 (2006).

28
Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6, e107 (2008).

29
Schadt, E. E., Woo, S. & Hao, K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat. Genet.* 44, 603–8 (2012).

30
Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* 339, 321–4 (2013).

31
Piskol, R., Ramaswami, G. & Li, J. B. B. Reliable Identification of Genomic Variants from RNA-Seq Data. *Am. J. Hum. Genet.* 93, 641–51 (2013).

32
Qin, S., Kim, J., Arafat, D. & Gibson, G. Effect of normalization on statistical and biological interpretation of gene expression profiles. *Front. Genet.* 3, 160 (2012).

33
Schurmann, C. *et al.* Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PLoS One* 7, e50938 (2012).

34
Mostafavi, S. *et al.* Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One* 8, e68141 (2013).

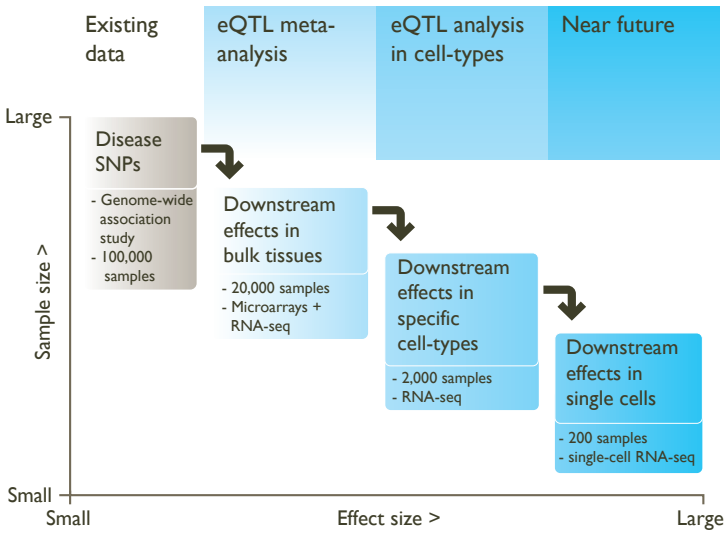


Figure 1.

Future eQTL mapping studies will likely focus on: 1) increasing the sample size through meta-analysis, in order to find more small-effect size eQTLs, 2) increasing the number of available tissues and cell types, in order to find cell-type-specific effects that are larger, and 3) single cell sequencing in order to identify context-specific eQTLs that have even larger effect sizes.

reliably. In Chapter 5, we focused on eQTLs derived from long intergenic non-coding transcripts (lincRNAs). We showed that the genetic regulation of some of these lincRNAs is tissue-specific, that different variants regulate non-coding and protein-coding transcripts, and that some trait-associated SNPs also regulate non-coding transcripts. However, we performed this study using microarrays that were designed before many non-coding transcripts were known. As such, we could study only a limited number of lincRNAs. Studies that apply RNA-seq to measure gene expression do not have this limitation and these will likely be able to detect many more lincRNA eQTLs in the near future.

RNA-seq is not limited to the measurement of whole transcript abundance: there are special protocols to measure the gene expression at certain ends of the transcripts: for example, GRO-seq can be used to measure nascent transcription near the 5' end of genes, while DeepSAGE has been developed to specifically measure gene expression near the 3' end, which is often the binding site for non-coding regulatory RNAs. We applied the DeepSAGE method in Chapter 6 and observed that RNA-seq data is more powerful for detecting eQTLs than microarray data due to its higher signal-to-noise ratios. Additionally, DeepSAGE can be used to assess the length of the poly-A tail of transcripts. We found 12 instances where a genetic variant influenced the length of a particular transcript, indicating that SNPs can regulate gene expression in various ways. However, the sample size of this study was rather small, which means that many small-effect eQTLs (and especially *trans*-eQTLs) have likely been missed. To overcome this we performed a meta-analysis of eQTL datasets generated using different RNA-seq protocols. By integrating different types of eQTL datasets, we obtained a comprehensive overview of the full spectrum of transcription regulation, including the differences in overall gene expression levels, isoform ratios, poly-A tail usage,

transcription-start efficiency, and exon usage.

Finally, we performed two large-scale eQTL meta-analyses (Chapters 3 and 7), in which we performed a *cis*- and *trans*-eQTL meta-analysis (initially on 1,469 individuals from two datasets, and subsequently on 5,311 individuals from nine datasets). We showed that, in whole blood, *trans*-eQTLs can be reliably detected and also replicated in independent datasets. We showed that certain SNPs affect the expression levels of many genes that operate within the same pathway. We demonstrated, to our knowledge for the first time, that independent SNPs that cause the same disease can affect exactly the same downstream genes in trans. For some disease-associated SNPs, the downstream *trans*-gene expression effects reflect the known hallmarks of these diseases, even though we had conducted these meta-analyses in healthy individuals. Additionally, we showed that trait-associated SNPs are enriched for both *cis*- and *trans*-eQTLs and that reported GWAS associations with a high significance are more likely to result in eQTL effects than associations that were less significant. These results indicate that eQTL mapping can reveal the downstream effects of disease SNPs, which may be helpful in the development of drugs to target these downstream genes. However, our results were derived from whole blood data, while many of the trait-associated SNPs in the current GWAS catalog are immune-related. As such, the downstream effects of these SNPs can be picked up in whole blood. The question remains whether this is also the case for complex brain diseases, for example. Especially for *trans*-eQTLs, the tissue specificity remains to be investigated: the tissue-specific datasets that are currently available often have too few samples to detect *trans*-eQTL effects properly. However, RNA-seq data for many tissues will soon become available in sufficient sample sizes through repositories such as the European Nucleotide Archive: since RNA-sequence reads also provide information on SNP genotypes, it is possible to derive genotype data, enabling eQTL analysis. By re-using such RNA-seq data, originally generated by many different labs throughout the world for different research purposes, we expect it to be possible to identify *trans*-eQTLs in many different tissues in the near future.

Part 2 – Future perspectives

Although eQTL study sample sizes keep on increasing, sample sizes are still limited in their capacity to detect many *trans*-eQTLs and *cis*-eQTL effects for rare variants. Additionally, because the range of available eQTL studies on different cell types and tissues is limited, the question about the eQTL effects of trait-associated SNPs in different tissues remains largely unanswered. Future eQTL studies will therefore likely focus on three different levels in order to find the downstream effects of trait-associated variants (Figure 1):
1) They will increase the sample size by meta-analysis to identify small-effect *cis*- and *trans*-eQTLs in bulk tissues (such as blood).
2) Based on these eQTLs, future studies will then zoom into specific cell types where the eQTLs show larger effects.
3) They will likely improve the results by using single-cell RNA sequencing to identify a highly specific cell-type or specific context

When genetic association studies, including eQTL studies, detect a significantly associated variant, this does not necessarily mean that the variant identified is also the variant causing the phenotypic variation (e.g. the difference in gene expression levels). There are several factors that can obscure the true causal variant in association studies. The associated variant may be in linkage disequilibrium (LD) with other variants in the locus: LD describes the likelihood of observing two alleles on the same haplotype and with increasing LD between variants, it becomes increasingly difficult to distinguish causal variants. Additionally, there may not have been enough power to test low frequency variants (e.g. population-specific variants), or these variants may not have been included in the test because they were not present on the genotyping platform or in the imputation reference set. Meta-analysis can help to pinpoint the causal variant: by combining association signals from multiple independent datasets, statistical power can be increased to detect significant effects for low frequency variants, and to determine what is the most significantly associated variant for variants that are in strong (but not perfect) LD, thereby helping to pinpoint the causal variant.

in which these eQTLs show particularly strong effects (e.g. a specific context could be when a blood cell has been activated by a viral or bacterial trigger).

Allele-specific expression

Except for Chapter 6, all the studies in this paper have been performed on gene expression data obtained from microarrays. Because microarrays are only able to measure gene expression levels as an average over all alleles in diploid organisms, local eQTL effects were previously annotated as being *cis*-eQTL effects, while the traditional *cis* definition (Box 1) implies that gene expression originates from the same allele as the variant (i.e. allelic imbalance of transcription, or allele-specific expression, ASE). Previously, ASE was measured through RT-PCR, precluding genome-wide assessment of ASE. This can now be assessed genome-wide through RNA-seq: by assessing the number of reads in heterozygote individuals, and by inferring haplotypes from reference datasets, RNA-seq is able to determine ASE on a large scale (Figure 2). There is a relationship between eQTLs and ASE: eQTLs and ASE frequently overlap[35,36] and the number of reads in the ASE signal correlate with the effect size of the eQTL[17]. ASE without an overlapping eQTL signal suggests that the ASE variants are rare and ASE can thus be applied to detect *cis*-eQTL effects originating from rare variants[13,36]. The mechanisms behind ASE are still unclear: one study suggested that ASE is mediated by cis regulatory elements (CREs; elements such as DNAse-I hypersensitive regions, enhancers, etc.)[17], while another study suggested that ASE is genetic rather than epigenetic and may be mediated by transcript structure variation[13]. ASE can also be assessed as a quantitative trait, in order to map aseQTLs: 641 SNPs with an eQTL also showed an aseQTL, some of which were located more than 1 megabase away from the transcription start site[36].

Larger meta-analyses

Current eQTL studies, including those that have been presented in this thesis, have identified *cis*-eQTL effects for the majority of protein-coding genes. However, it is likely that *trans*-eQTL effects are even more numerous, although their effects are likely to be very small: of the 430 *trans*-eQTLs detected in the meta-analysis presented in Chapter 7, more than 70% explained

35
Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–72 (2010).
36
Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24, 14–24 (2014).

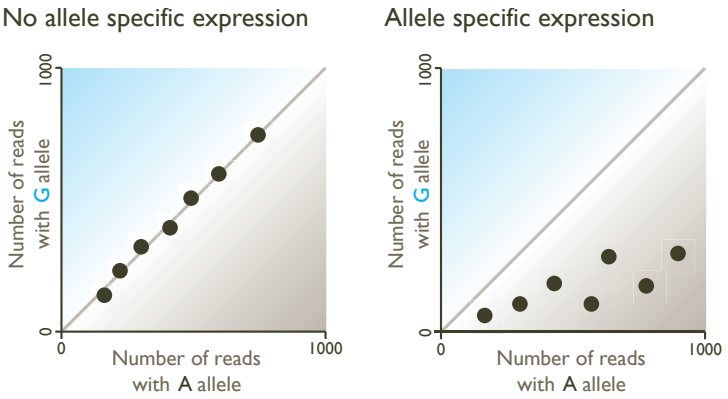


Figure 2.

Allele-specific expression can be determined using RNA-seq based gene expression data and phased genotypes. The number of reads is measured for individuals that are heterozygous for a certain exonic variant. If the number of reads per individual (represented as a dot in the figure) is more or less equal for both alleles (left), there is no allele-specific expression, and when there is a difference in the number of reads for a given individual (right), there is allele-specific expression.

less than 1% of the gene expression variation. Despite their small sample size, however, detecting more *trans*-eQTLs is important to determine the downstream effects of disease variants, especially when these *trans*-eQTL effects converge on the same genes, since this may help identify key disease driver genes (Figure 3). Thus, in order to find more *trans*-eQTLs with even smaller effect sizes, eQTL studies should be scaled up, similar to what was done for GWAS. Such meta-analyses will also permit us to fine-map existing eQTL more accurately, providing a higher-resolution overview of the downstream effects of both common and rare SNPs (Box 2). However, the cumulative sample size for many individual tissues or cell types is not yet large enough to find small-effect eQTLs. To overcome this, several methods have now been developed that allow meta-analysis over different tissues simultaneously[37,38]. Such large-scale, multi-tissue, eQTL meta-analyses will likely generate important biological insights into the downstream effects of many trait-associated variants.

An important issue that remains is multiple testing. When performing *trans*-eQTL analyses, billions of statistical tests need to be conducted. However, with ever increasing knowledge on the genes that are involved in specific pathways, it will also become possible to leverage external biological knowledge on these pathways to improve the statistical power: by averaging the expression levels of multiple genes that work in a specific pathway, signal-to-noise ratios can be improved, because noisy measurements of individual genes will be combined into a more robust and single pathway activity estimate. This will also result in fewer statistical tests being needed, because instead of testing SNPs against every human gene, only a limited number of pathways will need to be tested. Although such pathway-based eQTL methods have been proposed[39], few studies have used them so far in a human setting[25].

Finally, another important issue is to have access to the large amounts of eQTL datasets that have been produced so far.

37
Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* 9, e1003486 (2013).
38
Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* 9, e1003491 (2013).
39
Li, S., Lu, Q. & Cui, Y. A systems biology approach for identifying novel pathway regulators in eQTL mapping. *J. Biopharm. Stat.* 20, 373–400 (2010).

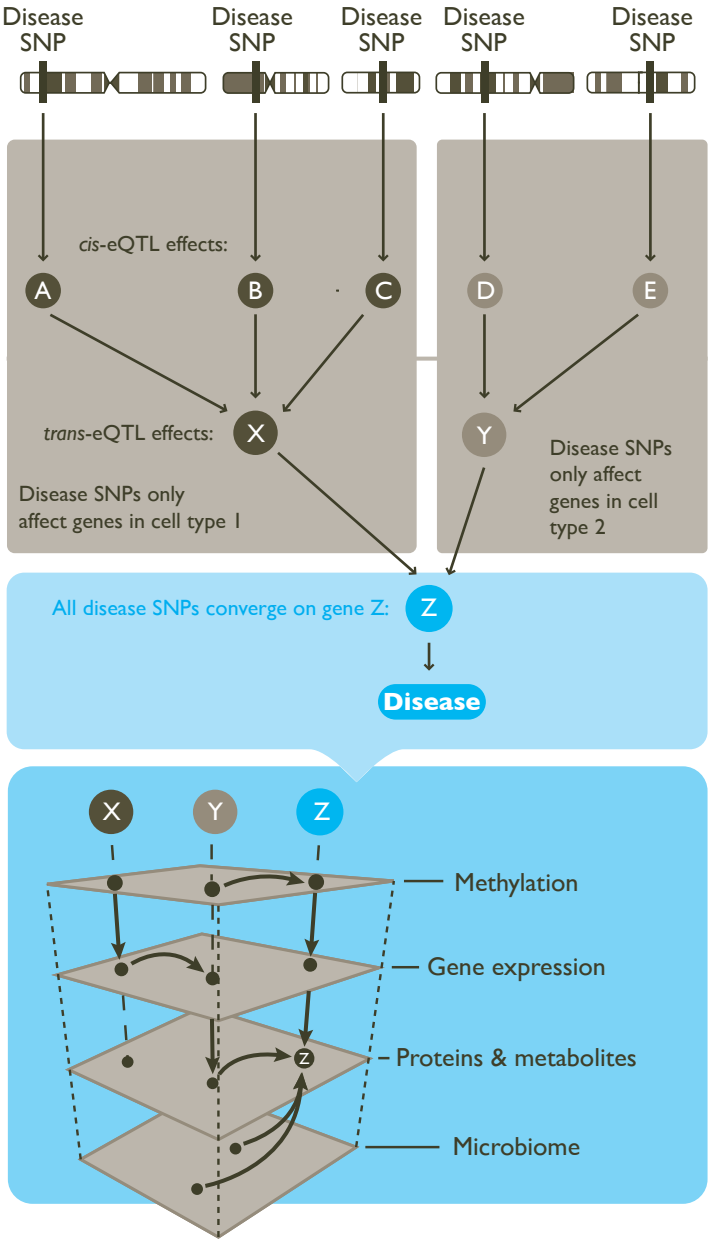


Figure 3.

Unraveling how disease-associated SNPs eventually cause disease will require an approach that integrates multiple phenotypes. First, using cell-type-specific eQTL datasets, we can define the effects of disease variants on the gene expression of local genes, or *cis*-eQTLs. These *cis*-eQTL effects (for example on a transcription factor) can also cause *trans*-eQTL effects, some of which may converge on the same gene, possibly across multiple tissues. We can thus consider these genes as key genes for the disease. Finally, different phenotypes, such as methylation levels, protein levels and the microbiome can be analyzed for genetic effects, interactions between the key disease genes, and the different phenotypic levels.

For instance, human whole blood eQTL data is now available for over 20,000 samples. Joint re-analysis of this data will likely yield important biological insight into the downstream effects of many trait-associated variants. Additionally, it is possible to use such whole blood eQTL datasets to make inferences about the specific cell-types in which eQTLs manifest themselves (by using the abundance of such cell types as an interaction term and by performing meta-analysis across different datasets).
However, while gene expression data is generally available (through databases such as the Gene Expression Omnibus, ArrayExpress, and more recently the European Nucleotide Archive), genotype data is available for only a limited number of eQTL datasets. This impedes progress on integrative approaches that can fully exploit such eQTL datasets to increase statistical power to identify smaller, but potentially very meaningful biological downstream eQTL effects. Initiatives such as dbGAP are therefore laudable, because they provide ways of sharing raw genotype data in a controlled and secure manner[30]. This might help to convince more researchers to make their data available to others (although researchers need to ensure they have ethical approval and informed consent from their patients that raw genotype data can be exchanged). When (legal) hurdles preclude such sharing, alternative strategies, such as performing eQTL meta-analyses[40] (where no raw genotype data, but only summary statistics are being exchanged), might provide ways to share such data for gaining novel biological insight.

Larger tissue- and cell-type-specific datasets

Although current studies, including the studies presented in Chapters 4 and 5 of this thesis, have shown that numerous trait-associated variants act in a tissue- or cell-type-specific way, it is likely that many cell-type-specific eQTLs have been missed, because their relevant tissue or cell type has not been investigated so far. As such, the question what is the disrupted tissue or cell type remains unanswered in many diseases; it may be important to know this in order to better understand the molecular disease mechanism (Figure 3). To provide insight into this issue, large-scale studies are currently underway that interrogate many different tissues or cell types from the same individuals.
The Genotype-Tissue Expression project (GTEx)[41], for example, aims to sample a range of tissues from approximately 900 samples. Gene expression is quantified using RNA-seq, which will enable the GTEx project to answer questions about the tissue specificity of *cis*- and *trans*-eQTLs, but will also provide insight into transcript isoform differences, ASE and differential exon usage between tissues. One of the aims of the GTEx project is to sample similar tissues to those used in the ENCODE project, which will then add information about the tissue-specific epigenetic signals underlying the regulation of gene expression caused by genetic variants (e.g. DNASe-I hypersensitivity and various histone modifications). Although the GTEx project will provide an extremely valuable resource for identifying the tissue specificity of eQTLs, each of these tissues still consists of many different cell types. Any of these cell types might be responsible for the observed eQTL effect in the compound tissue[5,9,42,43]. To detect these cell-type-specific eQTLs in blood, the ongoing ImmVar project[44] is focusing on a number of purified immunological cell types from approximately 600 individuals. This dataset will later be extended by an additional 28

40
Westra, H.-J. *et al.* Systematic identification of trans-eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–43 (2013).
41
Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–5 (2013).
42
Gregersen, P. K. Cell type–specific eQTLs in the human immune system. *Nat. Genet.* 44, 478–480 (2012).
43
Gerrits, A. *et al.* Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet.* 5, e1000692 (2009).
44
Lee, M. N. H. M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* (80-.). 343, 1246980–1246980 (2014).

cell types. With the full dataset, the ImmVar project will likely provide insight into the downstream effects of many genetic variants associated with immune-related diseases.

Single-cell eQTL analysis

All of the studies presented in this thesis, and nearly all other published eQTL studies, have so far been conducted while interrogating multiple cells per sample simultaneously. Since many eQTL papers have moved from studying whole tissues to individual cell types, the next logical step would be to study individual cells. This is particularly interesting, because it will permit identification of eQTLs that could well depend on the specific context in which a cell operates. Some effects of genetic variants on gene expression levels might manifest themselves only in those cells that have just been activated by a certain external stimulus (e.g. viral, bacterial, or other relevant trigger). If disease-associated SNPs only work in such a context, it could well be that such effects are not detectable when studying cells in bulk. The first paper to investigate this concept used single-cell RNA-seq and concentrated on 1,440 single cells from 15 individuals[45]. Many eQTLs were only detectable when studying single cells, and these would have been missed if the expression had been averaged over multiple cells.
Another attractive property of single-cell RNA-seq is that it might address a long-standing question in biology: how many different cell types are present in a given tissue (e.g. blood)? This can be determined by performing single-cell RNA-seq and generating expression profiles on thousands of individual cells, with subsequent comparison of the expression profiles (e.g. through principal component analysis). Although there are still many challenges on how to generate and analyze single-cell RNA seq data reliably and robustly[46,47], this technology will likely mature quickly, leading to much lower costs and permitting many research groups to initiate single-cell eQTL studies in the near future.

Molecular phenotypes underlying gene expression regulation

Although eQTLs provide insight into the downstream effects of genetic variation, they do not explain how gene expression is regulated by the genetic variant (Figure 3). Apart from regulation by other transcripts, gene expression may also be regulated by epigenetic marks: for example, some bases within the DNA, especially cytosine, can be methylated, which can modulate gene expression. Additionally, DNA is wrapped around complexes of histone proteins (histones consist of 4 subunits, H1 – H4), maintaining a higher order structure in the form of chromatin. Chromatin that is tightly packed with histones (heterochromatin) is less accessible to the transcription machinery and is thus associated with repression of transcription, while chromatin that is loosely packed with histones (euchromatin) is associated with active transcription. The density of chromatin packaging can be measured by treating the DNA-protein complex with DNAse-I (an enzyme that degrades DNA), followed by next generation sequencing (DNAse-seq) of the DNA fragments that were not digested. Treating the reads as a quantitative trait, a study on DNAse-I hypersensitivity QTLs (dsQTL) estimated that up to 55% of the eQTL SNPs were most likely also dsQTL SNPs[48].
Furthermore, the different histone proteins themselves can also be modified with, for example, methyl or acetyl groups, which can be

45 Wills, Q. F. *et al.* Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* 31, 748–52 (2013).

46 Ning, L. *et al.* Current Challenges in the Bioinformatics of Single Cell Genomics. *Front. Oncol.* 4, 7 (2014).

47 Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* 11, 22–24 (2013).

48 Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–4 (2012).

measured using ChIP-seq. These different modifications are associated with different processes, such as promotor activity and accessibility of chromatin. ChIP-seq data has been generated for many different tissues and cell lines for the ENCODE project and show enrichment for trait-associated SNPs within functional elements[49]. A recent study that assessed the overlap between trait-associated SNPs and such histone modifications reported that trait-associated SNPs often overlapped with specific histone modifications and that these signals were specific for certain cell types[50]. Finally, using chromatin conformation capture (3C and its variations 4C, 5C and Hi-C), interactions between loops of chromatin can be detected, which may also affect gene expression[51], and these loops have recently been linked to *trans*-eQTLs[52]. Although these epigenetic studies have been useful in elucidating the regulatory potential of genetic variants, the studies were generally performed using small sample sizes. As sample sizes increase, future studies will likely provide insight into the relationship of these different epigenetic signals and the *cis*- and *trans*-regulation of gene expression. The different epigenetic signals may also interact with each other to affect gene expression. Integrative analysis of the different epigenetic layers with gene expression will help to elucidate the precise mechanism of genetic regulation that determines the final gene expression.

Regulation of gene expression by other transcripts

Apart from protein-coding genes, a large fraction of the transcripts are non-coding and include piRNAs, snoRNAs, miRNAs and lncRNAs. piRNAs are transcripts of 26-31 nucleotides; they regulate the transcription of genes in germline cells by binding to PIWI proteins and may also alter DNA methylation levels of these genes[53,54]. snoRNAs are small ncRNAs of up to 300 nucleotides that function in the nucleolus, where they regulate ribosomal RNA stability[53]. miRNAs are small transcripts of 19-24 nucleotides that promote mRNA degradation or inhibit translation by binding through the RISC complex[53]. The specific binding sites of these miRNAs are hard to predict but they often appear to be located near the 3' UTR of transcripts[55]. lncRNAs are a family of long ncRNAs (>200 nucleotides), which may regulate gene expression through regulation of epigenetic signals or by promoting degradation or stabilization of mRNAs[53].The targets of many of these ncRNAs are still unknown, partly because probes for many of these transcripts, especially the small (< 50 bp) ncRNAs, are often not present on the microarray platforms. Additionally, for these small ncRNAs, special protocols for RNA-sequencing should be applied[56]. As a consequence, the expression of these transcripts has not yet been systematically assessed in large numbers of samples. However, these transcripts may provide additional mechanisms for *cis*- and *trans*-eQTLs: for example, variants in the 3' UTR of genes can affect the binding sites of miRNAs, leading to deregulation of the transcript abundance. Or, variants located in ncRNAs may alter the binding affinity with their target transcripts, causing deregulation and subsequent *trans* effects. In Chapter 5 of this thesis we showed that lincRNAs are often influenced by genetic variation. Future, large-scale, RNA-seq-based studies that investigate these non-coding transcripts will likely reveal that many of them are under genetic control.

49 Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).

50 Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124–30 (2013).

51 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* 295, 1306–11 (2002).

52 Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res.* 42, 1–10 (2013).

53 Esteller, M. Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–74 (2011).

54 Luteijn, M. J. & Ketting, R. F. PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nat. Rev. Genet.* 14, 523–34 (2013).

55 Pasquinelli, A. E. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat. Rev. Genet.* 13, 271–82 (2012).

56 Pritchard, C. C., Cheng, H. H. & Tewari, M. MicroRNA profiling: approaches and considerations. *Nat. Rev. Genet.* 13, 358–69 (2012).

Molecular phenotypes downstream of gene expression

eQTL studies have now provided a functional interpretation for many trait-associated SNPs. Future integrative approaches that also ascertain the effect of the trait-associated variants on different molecular levels (e.g. effects on protein levels, metabolite levels, or on the composition of the microbiome; Figure 3), and their possible interactions, will likely provide detailed mechanistic insights into the pathogenesis of many different diseases.

Appendices

A

Summary

DNA consists of 23 pairs of chromosomes which is present in almost every cell in the human body. A DNA molecule consists of two long intertwined strands of bases (which together form a double helix). The bases are chemical compounds that are abbreviated by the letters A, T, C and G and they encode genetic information. This genetic information can be transcribed into RNA molecules, which can have different functions. RNA molecules can regulate the transcription of other parts of the genome (transcription) or can be translated into a protein (translation). Depending on the origin of the cell (e.g. liver or brain cells), different parts of the genetic information can be transcribed, and as a consequence, different proteins will be produced, causing the cell to display different characteristics and functions in different organs or tissues.

Changes in the DNA can alter the structure or the amounts of RNA molecules that are being produced, which can eventually cause disease. Some diseases arise from a single change in the DNA (mutations that cause Mendelian diseases), while the majority is caused by different combination of many DNA changes (so-called complex diseases).

Through genome-wide association studies (GWAS), changes in the DNA sequence have been identified that increase or decrease disease risk for complex diseases. These changes are in the form of single nucleotide polymorphisms (SNPs) and involve a single base-pair. However, the majority of these SNPs are not located within genes that encode proteins and the number of genes which surround these SNPs is highly variable. Moreover, some of these SNPs are located in regions that do not contain any genes at all. As a consequence, after a GWAS study has been completed, the actual mechanism between the identified SNP and the disease is still often unclear.

Because the first step in the biology between SNP and disease is transcription into RNA, this process has been widely studied. The level at which DNA is transcribed into RNA can be measured using microarray platforms and RNA-sequencing. Subsequently, the genotypes of the SNP can be correlated with the transcription levels, to identify expression quantitative trait loci (eQTLs). Since eQTLs describe the effect of the SNP on transcription levels, they provide information about the downstream effects of SNPs (particularly disease-associated SNPs), and can thus provide information about the mechanism between the SNP and a certain disease. eQTLs can be local effects (*cis*-eQTL), where the SNP is located within a couple of hundred thousand bases from the affected RNA transcript, or they can be distant effects (*trans*-eQTL), where the SNP can be located on a completely different chromosome to the affected transcript.

This thesis presents methods to increase the statistical power to detect eQTLs with small effect size (part 1) and also presents specific eQTL association studies that provide novel biological insight into the downstream effects of trait-associated and other SNPs (part 2).

In Chapter 2, we present *MixupMapper*, a method that can be applied to eQTL datasets to detect accidental sample mix-ups that can occur on collection or in the laboratory. We applied the method to six publicly available datasets to assess the effects of sample mix-ups on the detection of eQTLs. Additionally, we performed simulations to determine the effect of sample mix-ups on our ability to explain the heritability of GWAS traits. We showed that, on average, 3% of the samples in public databases have been accidentally mixed-up, and that these mix-ups can have a large effect on the researchers' ability to detect small genetic effects. Consequently, just a few sample mix-ups in a database can have a large effect on our ability to explain the heritability of (disease) phenotypes.

In Chapter 3 we provide further methodology to increase the statistical power to detect eQTLs, by normalization of gene expression data through principal component analysis (PCA). Since PCA is able to capture variation in gene expression data that is not caused by genetic effects, correction for these components increases the statistical power to identify eQTLs. To further increase power, we performed a meta-analysis between two whole-blood eQTL datasets, enabling us to increase the sample size to identify small-effect eQTLs. Using both approaches, we were able to detect 340 *trans*-eQTL effects, and showed that independent trait-associated SNPs can show convergent effects on the same downstream genes, providing insight into the downstream pathways that are affected by these SNPs.

We applied a similar meta-analysis approach to seven independent, whole-blood eQTL datasets in Chapter 4. We used an interaction model that incorporates information on the differences in cell-type proportions across samples. In order to estimate these proportions, we developed a method that is able to predict these measurements from the gene expression data itself. We finally showed that the interaction effects identified describe cell-type-specific eQTLs, by replicating them in six independent but purified cell-type-specific eQTL datasets.

Not all genes code for proteins: a considerable fraction of all genes is transcribed into non-coding transcripts. In chapter 5, we describe an investigation into the genetic regulation of these transcripts in five different tissues, in which we showed that long intergenic non-coding RNAs (lincRNAs) are often genetically regulated as well. We also showed that the genetic regulation of these lincRNAs is often tissue-specific.

Although genes can vary their total expression levels, they can also generate different types of transcripts (i.e. isoforms). Transcripts can differ at their 3' end: they typically contain a tail of repeated adenosine bases (poly-A tail), but the length of this tail can differ, which can influence the translation of the gene into protein. RNA-sequencing data is well suited for identifying these isoform and poly-A tail differences. In Chapter 6 we describe using the RNA-sequencing technique DeepSAGE to investigate the genetic regulation of the length and usage of different poly-A tails of transcripts and identified 12 poly-adenylation QTLs.

In Chapter 7 we report how we applied our methods to a large-scale eQTL meta-analysis of 5,311 independent, whole-blood samples. We identified 1,513 *trans*-eQTL effects for trait-associated SNPs, indicating that eQTL mapping is a powerful tool for providing biological insight into the downstream effects of trait-associated and other SNPs. We discovered that two independent SNPs that show a *cis*-eQTL effect on the same gene (*IKZF1*) resulted in completely different *trans*-eQTL effects, and that disease-associated SNPs can sometimes have *trans*-eQTL effects on genes that describe the hallmark molecular features of these diseases.

In Chapter 8, we evaluate all our results and provide possible future directions for the biological interpretation of disease-associated genetic variants.

The main conclusions of this thesis are:

- Trait-associated SNPs are enriched for local (*cis*-) eQTL and distant (*trans*-) eQTL effects.
- *Cis*- and especially *trans*-eQTLs can provide valuable insight into the downstream effects of trait-associated SNPs.
- Multiple, independent, trait-associated SNPs can affect the same downstream (*trans*-eQTL) genes, which can identify important 'driver' genes
- *Trans*-eQTL effects can describe the hallmark processes of disease, even when the eQTL effect has been detected while studying only healthy individuals.
- eQTL effects are not limited to protein-coding genes, but can also be detected for non-coding genes. They can also influence the poly-A tail of transcripts.
- eQTL effects can be cell-type-specific. This cell-type specificity can be detected in gene expression data obtained from whole tissues, which consist of many different cell types.
- The power to detect eQTLs can be improved by meta-analysis and by correcting gene expression data for non-genetic factors and sample mix-ups.

Samenvatting

DNA bestaat uit 23 paren chromosomen, die aanwezig zijn in vrijwel iedere cel in het lichaam. Een DNA molecuul bestaat uit twee vervlochten strengen (die gezamenlijk de vorm hebben van een dubbele helix) en zijn opgebouwd uit zogenaamde basenparen (chemische verbindingen die afgekort worden met de letters A, T, C en G). Deze basen dragen de genetische informatie. Deze genetische informatie kan overgeschreven worden tot zogenaamde RNA (transcriptie). RNA moleculen kunnen de transcriptie van andere RNA moleculen beïnvloeden en kunnen vertaald worden naar eiwitten (translatie). Welk deel van het DNA wordt overgeschreven tot RNA, verschilt van cel tot cel (een levercel heeft bijvoorbeeld een ander transcriptie programma dan een brein cel) en hierdoor worden er verschillende eiwitten geproduceerd, waardoor verschillende cellen verschillende functies en eigenschappen hebben.

Veranderingen in het DNA kunnen gevolgen hebben voor de hoeveelheid transcriptie of de structuur van RNA moleculen. Zulke veranderingen kunnen leiden tot ziekte. Sommige ziektes ontstaan vanuit een enkele verandering in het DNA (mutaties die zogenaamde Mendeliaanse ziekten veroorzaken), terwijl verreweg de meeste ziekten (zogenaamde complexe ziekten) veroorzaakt worden door combinaties van verschillende veranderingen in het DNA.

Door het uitvoeren van genoom-wijde associatie studies (GWAS) zijn er de afgelopen jaren veel veranderingen in het DNA (in de vorm van enkele basenpaar veranderingen, genaamd SNPs) gevonden die het risico voor het ontwikkelen van complexe ziektes verhogen. Echter, het merendeel van deze SNPs ligt niet in genen die eiwitten beschrijven en liggen soms op plekken in het DNA waar helemaal geen genen aanwezig zijn. Als een gevolg hiervan is het na het uitvoeren van een GWAS vaak lastig gebleken om te bepalen via welk mechanisme de geassocieerde varianten uiteindelijk ziekte veroorzaken.

Om dit inzicht te verkrijgen, is er recentelijk veel aandacht besteed aan het onderzoeken van (variatie in) transcriptie, omdat dit de eerste moleculair biologische stap is van genetische variant naar uiteindelijke ziekte. De mate van transcriptie (expressie) kan worden gemeten met technieken als microarrays en RNA-sequencing. Door de expressie van het transcript te correleren met de genotypen van de SNP kunnen zo ‘expression quantitative trait loci’ (eQTL) geïdentificeerd worden. Deze eQTL beschrijven het effect van de variatie in het DNA op de transcriptie niveaus en beschrijven als gevolg daarvan de directe functionele gevolgen van de variant en daarmee inzicht in het ziekte proces. eQTL kunnen lokale effecten zijn (*cis*-eQTL), waarbij de afstand tussen de SNP en het transcript tot enkele honderdduizenden basenparen beperkt is. Deze afstand kan echter ook groot zijn (*trans*-eQTL), waarbij de SNP miljoenen basenparen verwijderd kan zijn van het transcript, of op een compleet ander chromosoom kan liggen.

Dit proefschrift beschrijft methodes die de statistische power vergroten om deze eQTLs te kunnen detecteren (deel I), maar presenteert ook specifieke eQTL studies die het biologisch

inzicht in de functionele gevolgen van ziekte geassocieerde SNPs hebben vergroot (deel 2).

In hoofdstuk 2 presenteren we *MixupMapper*, een methode die toegepast kan worden op eQTL datasets om sample verwisselingen te detecteren, die kunnen ontstaan bij het verzamelen en verwerken van samples in het laboratorium. We passen deze methode toe op zes publiekelijk beschikbare eQTL datasets om het effect van sample verwisselingen op de eQTL resultaten te laten zien. Verder voeren we simulaties uit om te bepalen wat het effect is van sample verwisselingen op de mogelijkheid om de erfelijkheid van ziektes te bepalen uit GWAS resultaten. We tonen aan dat gemiddeld 3% van de samples in de publieke datasets per ongeluk verwisseld zijn, en dat een beperkt aantal sample verwisselingen een groot effect kan hebben op het vermogen om eQTL effecten te detecteren.

In hoofdstuk 3 laten we zien dat de statistische power om eQTLs te vinden verder kan worden vergroot door het toepassen van principale component analyse (PCA) op de gen expressie data. De principale componenten kunnen niet-genetische variatie in gen expressie beschrijven, waardoor het corrigeren van de gen expressie data voor deze componenten de statistische power vergroot. Daarnaast beschrijft hoofdstuk 3 dat de statistische power verder kan worden vergroot door het uitvoeren van een meta-analyse. Door het toepassen van beide methodes op twee bloed eQTL datasets, vergroten we het aantal bekende *trans*-eQTL effecten tot 340. Verder laten we zien dat twee onafhankelijk met ziekte geassocieerde *trans*-eQTL SNPs effecten convergerende effecten kunnen hebben op dezelfde genen, waardoor regulatoire netwerken gevormd worden.

Eenzelfde meta-analyse strategie is toegepast in hoofdstuk 4 op zeven onafhankelijke eQTL datasets. In dit hoofdstuk passen we een interactiemodel toe, dat gebruik maakt van verschillen in celtype proporties in bloed tussen individuen, om de cel type specificiteit van eQTLs te detecteren. Om deze celtype proporties te schatten, beschrijven we een methode die in staat is de proporties te voorspellen aan de hand van gen expressie data. De voorspelde celtype specificiteit repliceren we vervolgens in zes onafhankelijke celtype specifieke eQTL datasets. Hiermee tonen we aan dat onze meta-analyse celtype specifieke eQTLs correct kan detecteren in een weefsel dat uit meerdere celtypen bestaat (in dit geval bloed).

Niet alle genen worden in eiwitten omgezet: een grote fractie van alle genen worden omgezet in zogenaamde non-coding RNAs (ncRNA). In hoofdstuk 5 onderzoeken we de genetische regulatie van deze transcripten in vijf verschillende weefsels, en laten zien dat bepaalde ncRNAs genetisch gereguleerd zijn en dat deze regulatie vaak weefselspecifiek is.

Genen worden niet alleen gereguleerd op het niveau van gen expressie. Zo kunnen er van een gen verschillende transcript varianten bestaan (zogenaamde isoformen). Daarnaast kan het 3' uiteinde van genen gewijzigd zijn: transcripten hebben vaak een staart die bestaat uit een herhaling van adenosine basen (poly-A staart). De lengte van deze staart kan verschillen, hetgeen de translatie naar het eiwit kan beïnvloeden. Deze verschillen in de

expressie van transcript isoformen en de poly-A staart lengte kunnen gedetecteerd worden met RNA-sequencing. In hoofdstuk 6 passen we de RNA-sequencing techniek DeepSAGE toe om de genetische regulatie van de poly-A staart lengte te onderzoeken en identificeren we 12 poly-adenlyatie QTLs.

In hoofdstuk 7 passen we de verschillende methodes om de statistische power te vergroten toe in een meta-analyse van in totaal 5,311 onafhankelijke bloed samples. Met deze meta-analyse identificeren we 1,513 *trans*-eQTL effecten voor met ziekte geassocieerde SNPs en tonen we aan dat eQTL associaties belangrijk inzicht kunnen leveren in de functionele gevolgen van deze varianten. Daarnaast laten we zien dat twee SNPs, die beide een *cis*-eQTL hebben op hetzelfde gen (*IKZF1*), verschillende *trans*-eQTL effecten kunnen hebben, en dat *trans*-eQTL effecten voor met ziekte geassocieerde varianten de kenmerkende eigenschappen van deze ziektes kunnen beschrijven.

In hoofdstuk 8 plaatsen we de resultaten van dit proefschrift in een bredere context en bespreken we nieuwe methoden die op termijn de biologische gevolgen van ziekte geassocieerde varianten kunnen help opsporen.

De belangrijkste conclusies van dit proefschrift zijn:

- Met ziekte geassocieerde SNPs zijn verrijkt voor *cis*- en *trans*-eQTL effecten.
- *Cis*- en vooral *trans*-eQTL effecten kunnen waardevolle inzichten bieden in de functionele gevolgen van met ziekte geassocieerde varianten.
- Meerdere, onafhankelijke, met ziekte geassocieerde varianten kunnen dezelfde *trans*-eQTL genen beïnvloeden en daarmee belangrijke, mogelijke 'driver' genen blootleggen.
- *Trans*-eQTL effecten kunnen de kenmerkende eigenschappen van ziektes beschrijven, zelfs als de eQTLs gedetecteerd zijn in gezonde individuen.
- eQTL effecten zijn niet beperkt tot genen die coderen voor eiwitten, maar kunnen ook gevonden worden voor genen die niet coderen voor eiwitten.
- Genetische variatie kan een effect hebben op de lengte van de poly-A staart van transcripten.
- eQTL effecten kunnen celtype specifiek zijn. Deze celtype specificiteit kan gedetecteerd worden in gen expressie data uit weefsels die uit meerdere celtypen bestaan.
- De statistische power om eQTLs te detecteren kan verhoogd worden door meta-analyse en door gen expressie data te corrigeren voor niet-genetische variatie en sample verwisselingen.

Curriculum vitae

Harm-Jan Westra werd geboren op 23 april 1984 te Beetgumermolen. Na het behalen van zijn diploma in 2002 op het Christelijk Gymnasium Beyers Naudé te Leeuwarden, studeerde hij een jaar lang medische beeldvormende technieken aan de Hanze Hogeschool te Groningen. In 2003 begon hij zijn bachelor opleiding Life Science & Technology aan de Rijksuniversiteit Groningen. Het diploma voor deze opleiding ontving hij in 2007, met als richting moleculaire en medische celbiologie. Hij vervolgde zijn studie met een masteropleiding bioinformatica aan het Wageningen Universiteit en Researchcentrum, waarvoor hij in 2009 zijn diploma ontving en als eindproject de gen expressie netwerken van een melkzuurbacterie bestudeerde. Deze opleiding stelde hem in staat om zijn interesse in zowel biologie als computertechniek te combineren. Begin 2010 begon hij als aio aan de afdeling genetica van het Universitair Medisch Centrum Groningen, onder supervisie van Dr. Lude Franke. Hij bestudeerde hoe genetische variatie gen expressie beïnvloedt wat uiteindelijk tot dit proefschrift heeft geleid. Hij zal zijn onderzoek voortzetten in de Verenigde Staten, aan het lab van Prof dr. Soumya Raychaudhuri, onderdeel van het Brigham and Women's Hospital en de Harvard Medical School te Boston.

Publicatielijst

33

Gockel, Ines, Becker, Jessica, Wouters, Mira M., Niebisch, Stefan, Gockel, Henning R., Hess, Timo, Ramonet, David, Zimmermann, Julian, Vigo, Ana González, Trynka, Gosia, de León, Antonio Ruiz, de la Serna, Julio Pérez, Urcelay, Elena, Kumar, Vinod, Franke, Lude, **Westra, Harm-Jan**, Drescher, Daniel, Kneist, Werner, Marquardt, Jens U., Galle, Peter R., Mattheisen, Manuel, Annese, Vito, Latiano, Anna, Fumagalli, Uberto, Laghi, Luigi, Cuomo, Rosario, Sarnelli, Giovanni, Müller, Michaela, Eckardt, Alexander J., Tack, Jan, Hoffmann, Per, Herms, Stefan, Mangold, Elisabeth, Heilmann, Stefanie, Kiesslich, Ralf, von Rahden, Burkhard H. A., Allescher, Hans-Dieter, Schulz, Henning G., Wijmenga, Cisca, Heneka, Michael T., Lang, Hauke, Hopfner, Karl-Peter, Nöthen, Markus M., Boeckxstaens, Guy E., de Bakker, Paul I. W., Knapp, Michael & Schumacher, Johannes. Common variants in the HLA-DQ region confer susceptibility to idiopathic achalasia. *Nat. Genet.* (2014). doi:10.1038/ng.3029

32

Cozen, W., Timofeeva, M. N., Li, D., Diepstra, A., Hazelett, D., Delahaye-Sourdeix, M., Edlund, C. K., Franke, L., Rostgaard, K., Van Den Berg, D. J., Cortessis, V. K., Smedby, K. E., Glaser, S. L., **Westra, H. J.**, Robison, L. L., Mack, T. M., Ghesquieres, H., Hwang, A. E., Nieters, A., de Sanjose, S., Lightfoot, T., Becker, N., Maynadie, M., Foretova, L., Roman, E., Benavente, Y., Rand, K. A., Nathwani, B. N., Glimelius, B., Staines, A., Boffetta, P., Link, B. K., Kiemeney, L., Ansell, S. M., Bhatia, S., Strong, L. C., Galan, P., Vatten, L., Habermann, T. M., Duell, E. J., Lake, A., Veenstra, R. N., Visser, L., Liu, Y., Urayama, K. Y., Montgomery, D., Gaborieau, V., Weiss, L. M., Byrnes, G., Lathrop, M., Cocco, P., Best, T., Skol, A. D., Adami, H. O., Melbye, M., Cerhan, J. R., Gallagher, A., Taylor, G. M., Slager, S. L., Brennan, P., Coetzee, G. A., Conti, D. V, Onel, K., Jarrett, R. F., Hjalgrim, H., van den Berg, A. & McKay, J. D. A meta-analysis of Hodgkin lymphoma reveals 19p13.3 TCF3 as a novel susceptibility locus. *Nat. Commun.* 5, 3856 (2014).

31

Deelen, Patrick, Menelaou, Androniki, van Leeuwen, Elisabeth M., Kanterakis, Alexandros, van Dijk, Freerk, Medina-Gomez, Carolina, Francioli, Laurent C., Hottenga, Jouke Jan, Karssen, Lennart C., Estrada, Karol, Kreiner-Møller, Eskil, Rivadeneira, Fernando, van Setten, Jessica, Gutierrez-Achury, Javier, **Westra, Harm-Jan**, Franke, Lude, van Enckevort, David, Dijkstra, Martijn, Byelas, Heorhiy, van Duijn, Cornelia M., de Bakker, Paul I. W., Wijmenga, Cisca & Swertz, Morris A. Improved imputation quality of low-frequency and rare variants in European samples using the "Genome of The Netherlands." *Eur. J. Hum. Genet.* (2014). doi:10.1038/ejhg.2014.19

30

Li, Haiying, Chan, Lillienne, Bartuzi, Paulina, Melton, Shelby D., Weber, Axel, Ben-Shlomo, Shani, Varol, Chen, Raetz, Megan, Mao, Xicheng, Starokadomskyy, Petro, van Sommeren, Suzanne, Mokadem, Mohamad, Schneider, Heike, Weisberg, Reid, **Westra, Harm-Jan**, Esko, Tõnu, Metspalu, Andres, Kumar, Vinod, Faubion, William A., Yarovinsky, Felix, Hofker, Marten, Wijmenga, Cisca, Kracht, Michael, Franke, Lude, Aguirre, Vincent, Weersma, Rinse K., Gluck, Nathan, van de Sluis, Bart & Burstein, Ezra. Copper metabolism domain-containing 1 represses genes that promote inflammation and protects mice from colitis and colitis-associated cancer. *Gastroenterology* 147, 184–195.e3 (2014).

29

Westra, Harm-Jan & Franke, Lude. From genome to function by studying eQTLs. *Biochim. Biophys. Acta* (2014). doi:10.1016/j.bbadis.2014.04.024

28

Fransen K, van Sommeren S, **Westra HJ**, Veenstra M, Lamberts LE, Modderman R, Dijkstra G, Fu J, Wijmenga C, Franke L, Weersma RK, van Diemen CC. Correlation of Genetic Risk and Messenger RNA Expression in a Th17/IL23 Pathway Analysis in Inflammatory Bowel Disease. *Inflamm Bowel Dis.* 2014 Mar 20.

27

Westra, H. J., Arends, D., Esko, T., Peters, M. J., Schurmann, C., Schramm, K., Kettunen, J., Yaghootkar, H., Fairfax, B., Andiappan, A. K., Li, Y., Fu, J., Karjalainen, J., Platteel, M., Visschedijk, M., Weersma, R., Kasela, S., Milani, L., Tserel, L., Peterson, P., Reinmaa, E., Hofman, A., Uitterlinden, A. G., Rivadeneira, F., Homuth, G., Petersmann, A., Lorbeer, R., Prokisch, H., Meitinger, T., Herder, C., Roden, M., Grallert, H., Ripatti, S., Perola, M., Wood, A. R., Melzer, D., Ferrucci, L., Singleton, A. B., Hernandez, D. G., Knight, J. C., Melchiotti, R., Lee, B., Poidinger, M., Zolezzi, F., Larbi, A., Wang, D. Y., van den Berg, L. H., Veldink, J. H., Rotzschke, O., Makino, S., Frayling, T., Salomaa, V., Strauch, K., Volker, U., van Meurs, J. B. J., Metspalu, A., Wijmenga, C., Jansen, R. C. & Franke, L. Cell specific eQTL analysis without sorting cells. *bioRxiv* (Cold Spring Harbor Labs Journals, 2014). doi:10.1101/002600

26

Hemani, Gibran, Shakhbazov, Konstantin, **Westra, Harm-Jan**, Esko, Tonu, Henders, Anjali K., McRae, Allan F., Yang, Jian, Gibson, Greg, Martin, Nicholas G., Metspalu, Andres, Franke, Lude, Montgomery, Grant W., Visscher, Peter M. & Powell, Joseph E. Detection and replication of epistasis influencing transcription in humans. *Nature* advance on, (2014).

25

Kumar, Vinod, **Westra, Harm-Jan**, Karjalainen, Juha, Zhernakova, Daria V, Esko, Tõnu, Hrdlickova, Barbara, Almeida, Rodrigo, Zhernakova, Alexandra, Reinmaa, Eva, Võsa, Urmo, Hofker, Marten H., Fehrmann, Rudolf S. N., Fu, Jingyuan, Withoff, Sebo, Metspalu, Andres, Franke, Lude & Wijmenga, Cisca. Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. *PLoS Genet.* 9, e1003201 (2013).

24

Cvejic, Ana, Haer-Wigman, Lonneke, Stephens, Jonathan C., Kostadima, Myrto, Smethurst, Peter A., Frontini, Mattia, Van Den Akker, Emile, Bertone, Paul, Bielczyk-Maczyńska, Ewa, Farrow, Samantha, Fehrmann, Rudolf S. N., Gray, Alan, De Haas, Masja, Haver, Vincent G., Jordan, Gregory, Karjalainen, Juha, Kerstens, Hindrik H. D., Kiddle, Graham, Lloyd-Jones, Heather, Needs, Malcolm, Poole, Joyce, Soussan, Aicha Ait, Rendon, Augusto, Rieneck, Klaus, Sambrook, Jennifer G., Schepers, Hein, Silljé, Herman H. W., Sipos, Botond, Swinkels, Dorine, Tamuri, Asif U., Verweij, Niek, Watkins, Nicholas A., **Westra, Harm-Jan**, Stemple, Derek, Franke, Lude, Soranzo, Nicole, Stunnenberg, Hendrik G., Goldman, Nick, Van Der Harst, Pim, Van Der Schoot, C. Ellen, Ouwehand, Willem H. & Albers, Cornelis A. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* 45, 542–5 (2013).

23

Rietveld, Cornelius A., Medland, Sarah E., Derringer, Jaime, Yang, Jian, Esko, Tõnu, Martin, Nicolas W., **Westra, Harm-Jan**, Shakhbazov, Konstantin, Abdellaoui, Abdel, Agrawal, Arpana, Albrecht, Eva, Alizadeh, Behrooz Z., Amin, Najaf, Barnard, John, Baumeister, Sebastian E., Benke, Kelly S., Bielak, Lawrence F., Boatman, Jeffrey A., Boyle, Patricia A., Davies, Gail, de Leeuw, Christiaan, Eklund, Niina, Evans, Daniel S., Ferhmann, Rudolf, Fischer, Krista, Gieger, Christian, Gjessing, Håkon K., Hägg, Sara, Harris, Jennifer R., Hayward, Caroline, Holzapfel, Christina, Ibrahim-Verbaas, Carla A., Ingelsson, Erik, Jacobsson, Bo, Joshi, Peter K., Jugessur, Astanand, Kaakinen, Marika, Kanoni, Stavroula, Karjalainen, Juha, Kolcic, Ivana, Kristiansson, Kati, Kutalik, Zoltán, Lahti, Jari, Lee, Sang H., Lin, Peng, Lind, Penelope A., Liu, Yongmei, Lohman, Kurt, Loitfelder, Marisa, McMahon, George, Vidal, Pedro Marques, Meirelles, Osorio, Milani, Lili, Myhre, Ronny, Nuotio, Marja-Liisa, Oldmeadow, Christopher J., Petrovic, Katja E., Peyrot, Wouter J., Polasek, Ozren, Quaye, Lydia, Reinmaa, Eva, Rice, John P., Rizzi, Thais S., Schmidt, Helena, Schmidt, Reinhold, Smith, Albert V, Smith, Jennifer A., Tanaka, Toshiko, Terracciano, Antonio, van der Loos, Matthijs J. H. M., Vitart, Veronique, Völzke, Henry, Wellmann, Jürgen, Yu, Lei, Zhao, Wei, Allik, Jüri, Attia, John R., Bandinelli, Stefania, Bastardot, François, Beauchamp, Jonathan, Bennett, David A., Berger, Klaus, Bierut, Laura J., Boomsma, Dorret I., Bültmann, Ute, Campbell, Harry, Chabris, Christopher F., Cherkas, Lynn, Chung, Mina K., Cucca, Francesco, de Andrade, Mariza, De Jager, Philip L., De Neve, Jan-Emmanuel, Deary, Ian J., Dedoussis, George V, Deloukas, Panos, Dimitriou, Maria, Eiríksdóttir, Guðny, Elderson, Martin F., Eriksson, Johan G., Evans, David M., Faul, Jessica D., Ferrucci, Luigi, Garcia, Melissa E., Grönberg, Henrik, Guðnason, Vilmundur, Hall, Per, Harris, Juliette M., Harris, Tamara B., Hastie, Nicholas D., Heath, Andrew C., Hernandez, Dena G., Hoffmann, Wolfgang, Hofman, Adriaan, Holle, Rolf, Holliday, Elizabeth G., Hottenga, Jouke-Jan, Iacono, William G., Illig, Thomas, Järvelin, Marjo-Riitta, Kähönen, Mika, Kaprio, Jaakko, Kirkpatrick, Robert M., Kowgier, Matthew, Latvala, Antti, Launer, Lenore J., Lawlor, Debbie A., Lehtimäki, Terho, Li, Jingmei, Lichtenstein, Paul, Lichtner, Peter, Liewald, David C., Madden, Pamela A., Magnusson, Patrik K. E., Mäkinen, Tomi E., Masala, Marco, McGue, Matt, Metspalu, Andres, Mielck, Andreas, Miller, Michael B., Montgomery, Grant W., Mukherjee, Sutapa, Nyholt, Dale R., Oostra, Ben A., Palmer, Lyle J., Palotie, Aarno, Penninx, Brenda W. J. H., Perola, Markus, Peyser, Patricia A., Preisig, Martin, Räikkönen, Katri, Raitakari, Olli T., Realo, Anu, Ring, Susan M., Ripatti, Samuli, Rivadeneira, Fernando, Rudan, Igor, Rustichini, Aldo, Salomaa, Veikko, Sarin, Antti-Pekka, Schlessinger, David, Scott, Rodney J., Snieder, Harold, St Pourcain, Beate, Starr, John M., Sul, Jae Hoon, Surakka, Ida, Svento, Rauli, Teumer, Alexander, Tiemeier, Henning, van Rooij, Frank J. A., Van Wagoner, David R., Vartiainen, Erkki, Viikari, Jorma, Vollenweider, Peter, Vonk, Judith M., Waeber, Gérard, Weir, David R., Wichmann, H. Erich, Widen, Elisabeth, Willemsen, Gonneke, Wilson, James F., Wright, Alan F., Conley, Dalton, Davey-Smith, George, Franke, Lude, Groenen, Patrick J. F., Hofman, Albert, Johannesson, Magnus, Kardia, Sharon L. R., Krueger, Robert F., Laibson, David, Martin, Nicholas G., Meyer, Michelle N., Posthuma, Danielle, Thurik, A. Roy, Timpson, Nicholas J., Uitterlinden, André G., van Duijn, Cornelia

M., Visscher, Peter M., Benjamin, Daniel J., Cesarini, David & Koellinger, Philipp D. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 340, 1467–71 (2013).

22

Den Hoed, Marcel, Eijgelsheim, Mark, Esko, Tõnu, Brundel, Bianca J. J. M., Peal, David S., Evans, David M., Nolte, Ilja M., Segrè, Ayellet V, Holm, Hilma, Handsaker, Robert E., **Westra, Harm-Jan**, Johnson, Toby, Isaacs, Aaron, Yang, Jian, Lundby, Alicia, Zhao, Jing Hua, Kim, Young Jin, Go, Min Jin, Almgren, Peter, Bochud, Murielle, Boucher, Gabrielle, Cornelis, Marilyn C., Gudbjartsson, Daniel, Hadley, David, van der Harst, Pim, Hayward, Caroline, den Heijer, Martin, Igl, Wilmar, Jackson, Anne U., Kutalik, Zoltán, Luan, Jian'an, Kemp, John P., Kristiansson, Kati, Ladenvall, Claes, Lorentzon, Mattias, Montasser, May E., Njajou, Omer T., O'Reilly, Paul F., Padmanabhan, Sandosh, St Pourcain, Beate, Rankinen, Tuomo, Salo, Perttu, Tanaka, Toshiko, Timpson, Nicholas J., Vitart, Veronique, Waite, Lindsay, Wheeler, William, Zhang, Weihua, Draisma, Harmen H. M., Feitosa, Mary F., Kerr, Kathleen F., Lind, Penelope A., Mihailov, Evelin, Onland-Moret, N. Charlotte, Song, Ci, Weedon, Michael N., Xie, Weijia, Yengo, Loic, Absher, Devin, Albert, Christine M., Alonso, Alvaro, Arking, Dan E., de Bakker, Paul I. W., Balkau, Beverley, Barlassina, Cristina, Benaglio, Paola, Bis, Joshua C., Bouatia-Naji, Nabila, Brage, Søren, Chanock, Stephen J., Chines, Peter S., Chung, Mina, Darbar, Dawood, Dina, Christian, Dörr, Marcus, Elliott, Paul, Felix, Stephan B., Fischer, Krista, Fuchsberger, Christian, de Geus, Eco J. C., Goyette, Philippe, Gudnason, Vilmundur, Harris, Tamara B., Hartikainen, Anna-Liisa, Havulinna, Aki S., Heckbert, Susan R., Hicks, Andrew A., Hofman, Albert, Holewijn, Suzanne, Hoogstra-Berends, Femke, Hottenga, Jouke-Jan, Jensen, Majken K., Johansson, Asa, Junttila, Juhani, Kääb, Stefan, Kanon, Bart, Ketkar, Shamika, Khaw, Kay-Tee, Knowles, Joshua W., Kooner, Angrad S., Kors, Jan A., Kumari, Meena, Milani, Lili, Laiho, Päivi, Lakatta, Edward G., Langenberg, Claudia, Leusink, Maarten, Liu, Yongmei, Luben, Robert N., Lunetta, Kathryn L., Lynch, Stacey N., Markus, Marcello R. P., Marques-Vidal, Pedro, Mateo Leach, Irene, McArdle, Wendy L., McCarroll, Steven A., Medland, Sarah E., Miller, Kathryn A., Montgomery, Grant W., Morrison, Alanna C., Müller-Nurasyid, Martina, Navarro, Pau, Nelis, Mari, O'Connell, Jeffrey R., O'Donnell, Christopher J., Ong, Ken K., Newman, Anne B., Peters, Annette, Polasek, Ozren, Pouta, Anneli, Pramstaller, Peter P., Psaty, Bruce M., Rao, Dabeeru C., Ring, Susan M., Rossin, Elizabeth J., Rudan, Diana, Sanna, Serena, Scott, Robert A., Sehmi, Jaban S., Sharp, Stephen, Shin, Jordan T., Singleton, Andrew B., Smith, Albert V, Soranzo, Nicole, Spector, Tim D., Stewart, Chip, Stringham, Heather M., Tarasov, Kirill V, Uitterlinden, André G., Vandenput, Liesbeth, Hwang, Shih-Jen, Whitfield, John B., Wijmenga, Cisca, Wild, Sarah H., Willemsen, Gonneke, Wilson, James F., Witteman, Jacqueline C. M., Wong, Andrew, Wong, Quenna, Jamshidi, Yalda, Zitting, Paavo, Boer, Jolanda M. A., Boomsma, Dorret I., Borecki, Ingrid B., van Duijn, Cornelia M., Ekelund, Ulf, Forouhi, Nita G., Froguel, Philippe, Hingorani, Aroon, Ingelsson, Erik, Kivimaki, Mika, Kronmal, Richard A., Kuh, Diana, Lind, Lars, Martin, Nicholas G., Oostra, Ben A., Pedersen, Nancy L., Quertermous, Thomas, Rotter, Jerome I., van der Schouw, Yvonne T., Verschuren, W. M. Monique, Walker, Mark, Albanes, Demetrius, Arnar, David O., Assimes, Themistocles L., Bandinelli, Stefania, Boehnke, Michael, de Boer, Rudolf A., Bouchard, Claude, Caulfield, W. L. Mark, Chambers, John C., Curhan, Gary, Cusi, Daniele, Eriksson, Johan, Ferrucci, Luigi, van Gilst, Wiek H., Glorioso, Nicola, de Graaf, Jacqueline, Groop, Leif, Gyllensten, Ulf, Hsueh, Wen-Chi, Hu, Frank B., Huikuri, Heikki V, Hunter, David J., Iribarren, Carlos, Isomaa, Bo, Jarvelin, Marjo-Riitta, Jula, Antti, Kähönen, Mika, Kiemeney, Lambertus A., van der Klauw, Melanie M., Kooner, Jaspal S., Kraft, Peter, Iacoviello, Licia, Lehtimäki, Terho, Lokki, Marja-Liisa L., Mitchell, Braxton D., Navis, Gerjan, Nieminen, Markku S., Ohlsson, Claes, Poulter, Neil R., Qi, Lu, Raitakari, Olli T., Rimm, Eric B., Rioux, John D., Rizzi, Federica, Rudan, Igor, Salomaa, Veikko, Sever, Peter S., Shields, Denis C., Shuldiner, Alan R., Sinisalo, Juha, Stanton, Alice V, Stolk, Ronald P., Strachan, David P., Tardif, Jean-Claude, Thorsteinsdottir, Unnur, Tuomilehto, Jaako, van Veldhuisen, Dirk J., Virtamo, Jarmo, Viikari, Jorma, Vollenweider, Peter, Waeber, Gérard, Widen, Elisabeth, Cho, Yoon Shin, Olsen, Jesper V, Visscher, Peter M., Willer, Cristen, Franke, Lude, Erdmann, Jeanette, Thompson, John R., Pfeufer, Arne, Sotoodehnia, Nona, Newton-Cheh, Christopher, Ellinor, Patrick T., Stricker, Bruno H. Ch, Metspalu, Andres, Perola, Markus, Beckmann, Jacques S., Smith, George Davey, Stefansson, Kari, Wareham, Nicholas J., Munroe, Patricia B., Sibon, Ody C. M., Milan, David J., Snieder, Harold, Samani, Nilesh J. & Loos, Ruth J. F. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat. Genet.* 45, 621–31 (2013).

21

Zhernakova, Daria V, de Klerk, Eleonora, **Westra, Harm-Jan**, Mastrokolias, Anastasios, Amini, Shoaib, Ariyurek, Yavuz, Jansen, Rick, Penninx, Brenda W., Hottenga, Jouke J., Willemsen, Gonneke, de Geus, Eco J., Boomsma, Dorret I., Veldink, Jan H., van den Berg, Leonard H., Wijmenga, Cisca, den Dunnen, Johan T., van Ommen, Gert-Jan B., 't Hoen, Peter A. C. & Franke, Lude. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* 9, e1003594 (2013).

20

Westra, Harm-Jan, Peters, Marjolein J., Esko, Tõnu, Yaghootkar, Hanieh, Schurmann, Claudia, Kettunen, Johannes, Christiansen, Mark W., Fairfax, Benjamin P., Schramm, Katharina, Powell, Joseph E., Zhernakova, Alexandra, Zhernakova, Daria V, Veldink, Jan H., Van den Berg, Leonard H., Karjalainen, Juha, Withoff, Sebo, Uitterlinden, André G., Hofman, Albert, Rivadeneira, Fernando, 't Hoen, Peter A. C., Reinmaa, Eva, Fischer, Krista, Nelis, Mari, Milani, Lili, Melzer, David, Ferrucci, Luigi, Singleton, Andrew B., Hernandez, Dena G., Nalls, Michael A., Homuth, Georg, Nauck, Matthias, Radke, Dörte, Völker, Uwe, Perola, Markus, Salomaa, Veikko, Brody, Jennifer, Suchy-Dicey, Astrid, Gharib, Sina A., Enquobahrie, Daniel A., Lumley, Thomas, Montgomery, Grant W., Makino, Seiko, Prokisch, Holger, Herder, Christian, Roden, Michael, Grallert, Harald, Meitinger, Thomas, Strauch, Konstantin, Li, Yang, Jansen, Ritsert C., Visscher, Peter M., Knight, Julian C., Psaty, Bruce M., Ripatti, Samuli, Teumer, Alexander, Frayling, Timothy M., Metspalu, Andres, van Meurs, Joyce B. J. & Franke, Lude. Systematic

identification of trans-eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–43 (2013).

19

Trouw, L. A., Daha, N., Kurreeman, F. A. S., Böhringer, S., Goulielmos, G. N., **Westra, H. J.**, Zhernakova, A., Franke, L., Stahl, E. A., Levarht, E. W. N., Stoeken-Rijsbergen, G., Verduijn, W., Roos, A., Li, Y., Houwing-Duistermaat, J. J., Huizinga, T. W. J. & Toes, R. E. M. Genetic variants in the region of the C1q genes are associated with rheumatoid arthritis. *Clin. Exp. Immunol.* 173, 76–83 (2013).

18

Paul, Dirk S., Albers, Cornelis A., Rendon, Augusto, Voss, Katrin, Stephens, Jonathan, van der Harst, Pim, Chambers, John C., Soranzo, Nicole, Ouwehand, Willem H. & Deloukas, Panos. Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res.* 23, 1130–41 (2013).

17

Knevel, R., Krabben, A., Wilson, A. G., Brouwer, E., Leijsma, M. K., Lindqvist, E., de Rooy, D. P. C., Daha, N. A., van der Linden, M. P. M., Tsonaka, S., Zhernakova, A., **Westra, H. J.**, Franke, L., Houwing-Duistermaat, J. J., Toes, R. E. M., Huizinga, T. W. J., Saxne, T. & van der Helm-van Mil, A. H. M. A genetic variant in granzyme B is associated with progression of joint destruction in rheumatoid arthritis. Arthritis Rheum. 65, 582–9 (2013).

16

Almeida, Rodrigo, Ricaño-Ponce, Isis, Kumar, Vinod, Deelen, Patrick, Szperl, Agata, Trynka, Gosia, Gutierrez-Achury, Javier, Kanterakis, Alexandros, **Westra, Harm-Jan**, Franke, Lude, Swertz, Morris A., Platteel, Mathieu, Bilbao, Jose Ramon, Barisani, Donatella, Greco, Luigi, Mearin, Luisa, Wolters, Victorien M., Mulder, Chris, Mazzilli, Maria Cristina, Sood, Ajit, Cukrowska, Bozena, Núñez, Concepción, Pratesi, Riccardo, Withoff, Sebo & Wijmenga, Cisca. Fine mapping of the celiac disease-associated LPP locus reveals a potential functional variant. *Hum. Mol. Genet.* (2013). doi:10.1093/hmg/ddt619

15

Okada, Yukinori, Wu, Di, Trynka, Gosia, Raj, Towfique, Terao, Chikashi, Ikari, Katsunori, Kochi, Yuta, Ohmura, Koichiro, Suzuki, Akari, Yoshida, Shinji, Graham, Robert R., Manoharan, Arun, Ortmann, Ward, Bhangale, Tushar, Denny, Joshua C., Carroll, Robert J., Eyler, Anne E., Greenberg, Jeffrey D., Kremer, Joel M., Pappas, Dimitrios A., Jiang, Lei, Yin, Jian, Ye, Lingying, Su, Dinghong Ding-Feng, Yang, Jian, Xie, Gang, Keystone, Ed, **Westra, Harm-Jan**, Esko, Tõnu, Metspalu, Andres, Zhou, Xuezhong, Gupta, Namrata, Mirel, Daniel, Stahl, Eli A., Diogo, Dorothée, Cui, Jing, Liao, Katherine, Guo, Michael H., Myouzen, Keiko, Kawaguchi, Takahisa, Coenen, Marieke J. H. H., van Riel, Piet L. C. M. C. M., van de Laar, Mart A. F. J. F. J., Guchelaar, Henk-Jan, Huizinga, Tom W. J. J., Dieudé, Philippe, Mariette, Xavier, Bridges Jr., S. Louis, Zhernakova, Alexandra, Toes, Rene E. M. M., Tak, Paul P., Miceli-Richard, Corinne, Bang, So-Young, Lee, Hye-Soon, Martin, Javier, Gonzalez-Gay, Miguel A., Rodriguez-Rodriguez, Luis, Rantapää-Dahlqvist, Solbritt, Ärlestig, Lisbeth, Choi, Hyon K., Kamatani, Yoichiro, Galan, Pilar, Lathrop, Mark, Consortium, the RACI GARNET, Eyre, Steve, Bowes, John, Barton, Anne, de Vries, Niek, Moreland, Larry W., Criswell, Lindsey A., Karlson, Elizabeth W., Taniguchi, Atsuo, Yamada, Ryo, Kubo, Michiaki, Liu, Jun S., Bae, Sang-Cheol, Worthington, Jane, Padyukov, Leonid, Klareskog, Lars, Gregersen, Peter K., Raychaudhuri, Soumya, Stranger, Barbara E., De Jager, Philip L., Franke, Lude, Visscher, Peter M., Brown, Matthew A., Yamanaka, Hisashi, Mimori, Tsuneyo, Takahashi, Atsushi, Xu, Huji, Behrens, Timothy W., Siminovitch, Katherine A., Momohara, Shigeki, Matsuda, Fumihiko, Yamamoto, Kazuhiko, Plenge, Robert M. & Louis Bridges Jr, S. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381 (2013).

14

Fu, Jingyuan, Wolfs, Marcel G. M., Deelen, Patrick, **Westra, Harm-Jan**, Fehrmann, Rudolf S. N., Te Meerman, Gerard J., Buurman, Wim A., Rensen, Sander S. M., Groen, Harry J. M., Weersma, Rinse K., Van Den Berg, Leonard H., Veldink, Jan, Ophoff, Roel A., Snieder, Harold, Van Heel, David, Jansen, Ritsert C., Hofker, Marten H., Wijmenga, Cisca & Franke, Lude. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* 8, e1002431 (2012).

13

De Boer, Rudolf A., Verweij, Niek, Van Veldhuisen, Dirk J., **Westra, Harm-Jan**, Bakker, Stephan J. L., Gansevoort, Ron T., Muller Kobold, Anneke C., Van Gilst, Wiek H., Franke, Lude, Mateo Leach, Irene & Van Der Harst, Pim. A genome-wide association study of circulating galectin-3. *PLoS One* 7, e47385 (2012).

12

Van der Harst, Pim, Zhang, Weihua, Mateo Leach, Irene, Rendon, Augusto, Verweij, Niek, Sehmi, Joban, Paul, Dirk S., Elling, Ulrich, Allayee, Hooman, Li, Xinzhong, Radhakrishnan, Aparna, Tan, Sian-Tsung, Voss, Katrin, Weichenberger, Christian X., Albers, Cornelis A., Al-Hussani, Abtehale, Asselbergs, Folkert W., Ciullo, Marina, Danjou, Fabrice, Dina, Christian, Esko, Tõnu, Evans, David M., Franke, Lude, Gögele, Martin, Hartiala, Jaana, Hersch, Micha, Holm, Hilma, Hottenga, Jouke-Jan, Kanoni, Stavroula, Kleber, Marcus E., Lagou, Vasiliki, Langenberg, Claudia, Lopez, Lorna M., Lyytikäinen, Leo-Pekka, Melander, Olle, Murgia, Federico, Nolte, Ilja M., O'Reilly, Paul F., Padmanabhan, Sandosh, Parsa, Afshin, Pirastu, Nicola, Porcu, Eleonora, Portas, Laura, Prokopenko, Inga, Ried, Janina S., Shin, So-Youn, Tang, Clara S., Teumer, Alexander, Traglia, Michela, Ulivi, Sheila, **Westra, Harm-Jan**, Yang, Jian, Zhao, Jing Hua, Anni, Franco, Abdellaoui, Abdel, Attwood, Antony, Balkau, Beverley, Bandinelli, Stefania, Bastardot, François, Benyamin, Beben, Boehm, Bernhard O., Cookson, William O., Das, Debashish, de Bakker, Paul I. W., de Boer, Rudolf A., de Geus, Eco J. C., de Moor, Marleen H., Dimitriou, Maria, Domingues, Francisco S., Döring, Angela, Engström, Gunnar, Eyjolfsson, Gudmundur Ingi, Ferrucci, Luigi, Fischer, Krista, Galanello, Renzo, Garner, Stephen F., Genser, Bernd, Gibson, Quince D., Girotto, Giorgia, Gudbjartsson, Daniel Fannar, Harris, Sarah E., Hartikainen, Anna-Liisa, Hastie, Claire E., Hedblad, Bo, Illig, Thomas, Jolley, Jennifer, Kähönen, Mika, Kema, Ido P., Kemp, John P., Liang, Liming, Lloyd-Jones, Heather, Loos, Ruth J. F., Meacham, Stuart, Medland, Sarah E., Meisinger, Christa, Memari, Yasin, Mihailov, Evelin, Miller, Kathy, Moffatt, Miriam F., Nauck, Matthias, Novatchkova, Maria, Nutile, Teresa, Olafsson, Isleifur, Onundarson, Pall T., Parracciani, Debora, Penninx, Brenda W., Perseu, Lucia, Piga,

Antonio, Pistis, Giorgio, Pouta, Anneli, Puc, Ursula, Raitakari, Olli, Ring, Susan M., Robino, Antonietta, Ruggiero, Daniela, Ruokonen, Aimo, Saint-Pierre, Aude, Sala, Cinzia, Salumets, Andres, Sambrook, Jennifer, Schepers, Hein, Schmidt, Carsten Oliver, Silljé, Herman H. W., Sladek, Rob, Smit, Johannes H., Starr, John M., Stephens, Jonathan, Sulem, Patrick, Tanaka, Toshiko, Thorsteinsdottir, Unnur, Tragante, Vinicius, van Gilst, Wiek H., van Pelt, L. Joost, van Veldhuisen, Dirk J., Völker, Uwe, Whitfield, John B., Willemsen, Gonneke, Winkelmann, Bernhard R., Wirnsberger, Gerald, Algra, Ale, Cucca, Francesco, d'Adamo, Adamo Pio, Danesh, John, Deary, Ian J., Dominiczak, Anna F., Elliott, Paul, Fortina, Paolo, Froguel, Philippe, Gasparini, Paolo, Greinacher, Andreas, Hazen, Stanley L., Jarvelin, Marjo-Riitta, Khaw, Kay Tee, Lehtimäki, Terho, Maerz, Winfried, Martin, Nicholas G., Metspalu, Andres, Mitchell, Braxton D., Montgomery, Grant W., Moore, Carmel, Navis, Gerjan, Pirastu, Mario, Pramstaller, Peter P., Ramirez-Solis, Ramiro, Schadt, Eric, Scott, James, Shuldiner, Alan R., Smith, George Davey, Smith, J. Gustav, Snieder, Harold, Sorice, Rossella, Spector, Tim D., Stefansson, Kari, Stumvoll, Michael, Tang, W. H. Wilson, Toniolo, Daniela, Tönjes, Anke, Visscher, Peter M., Vollenweider, Peter, Wareham, Nicholas J., Wolffenbuttel, Bruce H. R., Boomsma, Dorret I., Beckmann, Jacques S., Dedoussis, George V, Deloukas, Panos, Ferreira, Manuel A., Sanna, Serena, Uda, Manuela, Hicks, Andrew A., Penninger, Josef Martin, Gieger, Christian, Kooner, Jaspal S., Ouwehand, Willem H., Soranzo, Nicole & Chambers, John C. Seventy-five genetic loci influencing the human red blood cell. *Nature* 492, 369–75 (2012).

11

Franceschini, Nora, van Rooij, Frank J. A., Prins, Bram P., Feitosa, Mary F., Karakas, Mahir, Eckfeldt, John H., Folsom, Aaron R., Kopp, Jeffrey, Vaez, Ahmad, Andrews, Jeanette S., Baumert, Jens, Boraska, Vesna, Broer, Linda, Hayward, Caroline, Ngwa, Julius S., Okada, Yukinori, Polasek, Ozren, **Westra, Harm-Jan**, Wang, Ying A., Del Greco M, Fabiola, Glazer, Nicole L., Kapur, Karen, Kema, Ido P., Lopez, Lorna M., Schillert, Arne, Smith, Albert V, Winkler, Cheryl A., Zgaga, Lina, Bandinelli, Stefania, Bergmann, Sven, Boban, Mladen, Bochud, Murielle, Chen, Y. D., Davies, Gail, Dehghan, Abbas, Ding, Jingzhong, Doering, Angela, Durda, J. Peter, Ferrucci, Luigi, Franco, Oscar H., Franke, Lude, Gunjaca, Grog, Hofman, Albert, Hsu, Fang-Chi, Kolcic, Ivana, Kraja, Aldi, Kubo, Michiaki, Lackner, Karl J., Launer, Lenore, Loehr, Laura R., Li, Guo, Meisinger, Christa, Nakamura, Yusuke, Schwienbacher, Christine, Starr, John M., Takahashi, Atsushi, Torlak, Vesela, Uitterlinden, André G., Vitart, Veronique, Waldenberger, Melanie, Wild, Philipp S., Kirin, Mirna, Zeller, Tanja, Zemunik, Tatijana, Zhang, Qunyuan, Ziegler, Andreas, Blankenberg, Stefan, Boerwinkle, Eric, Borecki, Ingrid B., Campbell, Harry, Deary, Ian J., Frayling, Timothy M., Gieger, Christian, Harris, Tamara B., Hicks, Andrew A., Koenig, Wolfgang, O' Donnell, Christopher J., Fox, Caroline S., Pramstaller, Peter P., Psaty, Bruce M., Reiner, Alex P., Rotter, Jerome I., Rudan, Igor, Snieder, Harold, Tanaka, Toshihiro, van Duijn, Cornelia M., Vollenweider, Peter, Waeber, Gerard, Wilson, James F., Witteman, Jacqueline C. M., Wolffenbuttel, Bruce H. R., Wright, Alan F., Wu, Qingyu, Liu, Yongmei, Jenny, Nancy S., North, Kari E., Felix, Janine F., Alizadeh, Behrooz Z., Cupples, L. Adrienne, Perry, John R. B. & Morris, Andrew P. Discovery and fine mapping of serum protein loci through transethnic meta-analysis. *Am. J. Hum. Genet.* 91, 744–53 (2012).

10

Eyre, Steve, Bowes, John, Diogo, Dorothée, Lee, Annette, Barton, Anne, Martin, Paul, Zhernakova, Alexandra, Stahl, Eli, Viatte, Sebastien, McAllister, Kate, Amos, Christopher I., Padyukov, Leonid, Toes, Rene E. M., Huizinga, Tom W. J., Wijmenga, Cisca, Trynka, Gosia, Franke, Lude, **Westra, Harm-Jan**, Alfredsson, Lars, Hu, Xinli, Sandor, Cynthia, de Bakker, Paul I. W., Davila, Sonia, Khor, Chiea Chuen, Heng, Khai Koon, Andrews, Robert, Edkins, Sarah, Hunt, Sarah E., Langford, Cordelia, Symmons, Deborah, Concannon, Pat, Onengut-Gumuscu, Suna, Rich, Stephen S., Deloukas, Panos, Gonzalez-Gay, Miguel A., Rodriguez-Rodriguez, Luis, Ärlsetig, Lisbeth, Martin, Javier, Rantapää-Dahlqvist, Solbritt, Plenge, Robert M., Raychaudhuri, Soumya, Klareskog, Lars, Gregersen, Peter K. & Worthington, Jane. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* 44, 1336–40 (2012).

9

Westra, Harm-Jan, Jansen, Ritsert C., Fehrmann, Rudolf S. N., Te Meerman, Gerard J., Van Heel, David, Wijmenga, Cisca & Franke, Lude. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics 27, 2104–2111 (2011).

8

Fehrmann, Rudolf S. N., Jansen, Ritsert C., Veldink, Jan H., **Westra, Harm-Jan**, Arends, Danny, Bonder, Marc Jan, Fu, Jingyuan, Deelen, Patrick, Groen, Harry J. M., Smolonska, Asia, Weersma, Rinse K., Hofstra, Robert M. W., Buurman, Wim A., Rensen, Sander, Wolfs, Marcel G. M., Platteel, Mathieu, Zhernakova, Alexandra, Elbers, Clara C., Festen, Eleanora M., Trynka, Gosia, Hofker, Marten H., Saris, Christiaan G. J., Ophoff, Roel A., Van Den Berg, Leonard H., van Heel, David A., Wijmenga, Cisca, Te Meerman, Gerard J. & Franke, Lude. *Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* 7, 14 (2011).

7

Janse, Marcel, Lamberts, Laetitia E., Franke, Lude, Raychaudhuri, Soumya, Ellinghaus, Eva, Muri Boberg, Kirsten, Melum, Espen, Folseraas, Trine, Schrumpf, Erik, Bergquist, Annika, Björnsson, Einar, Fu, Jingyuan, **Jan Westra, Harm**, Groen, Harry J. M., Fehrmann, Rudolf S. N., Smolonska, Joanna, Van Den Berg, Leonard H., Ophoff, Roel A., Porte, Robert J., Weismüller, Tobias J., Wedemeyer, Jochen, Schramm, Christoph, Sterneck, Martina, Günther, Rainer, Braun, Felix, Vermeire, Severine, Henckaerts, Liesbet, Wijmenga, Cisca, Ponsioen, Cyriel Y., Schreiber, Stefan, Karlsen, Tom H., Franke, Andre & Weersma, Rinse K. Three ulcerative colitis susceptibility loci are associated with primary sclerosing cholangitis and indicate a role for IL2, REL, and CARD9. *Hepatology* 53, 1977–85 (2011).

6

Zhernakova, Alexandra, Stahl, Eli A., Trynka, Gosia, Raychaudhuri, Soumya, Festen, Eleanora A., Franke, Lude, **Westra, Harm-Jan**, Fehrmann, Rudolf S. N., Kurreeman, Fina A. S., Thomson, Brian, Gupta, Namrata, Romanos, Jihane, McManus, Ross, Ryan, Anthony W., Turner, Graham, Brouwer, Elisabeth, Posthumus, Marcel D., Remmers, Elaine F., Tucci,

Francesca, Toes, Rene, Grandone, Elvira, Mazzilli, Maria Cristina, Rybak, Anna, Cukrowska, Bozena, Coenen, Marieke J. H., Radstake, Timothy R. D. J., Van Riel, Piet L. C. M., Li, Yonghong, De Bakker, Paul I. W., Gregersen, Peter K., Worthington, Jane, Siminovitch, Katherine A., Klareskog, Lars, Huizinga, Tom W. J., Wijmenga, Cisca & Plenge, Robert M. Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci. *PLoS Genet.* 7, 13 (2011).

5

Johnson, Toby, Gaunt, Tom R., Newhouse, Stephen J., Padmanabhan, Sandosh, Tomaszewski, Maciej, Kumari, Meena, Morris, Richard W., Tzoulaki, Ioanna, O'Brien, Eoin T., Poulter, Neil R., Sever, Peter, Shields, Denis C., Thom, Simon, Wannamethee, Sasiwarang G., Whincup, Peter H., Brown, Morris J., Connell, John M., Dobson, Richard J., Howard, Philip J., Mein, Charles A., Onipinla, Abiodun, Shaw-Hawkins, Sue, Zhang, Yun, Davey Smith, George, Day, Ian N. M., Lawlor, Debbie A., Goodall, Alison H., Fowkes, F. Gerald, Abecasis, Gonçalo R., Elliott, Paul, Gateva, Vesela, Braund, Peter S., Burton, Paul R., Nelson, Christopher P., Tobin, Martin D., van der Harst, Pim, Glorioso, Nicola, Neuvrith, Hani, Salvi, Erika, Staessen, Jan A., Stucchi, Andrea, Devos, Nabila, Jeunemaitre, Xavier, Plouin, Pierre-François, Tichet, Jean, Juhanson, Peeter, Org, Elin, Putku, Margus, Sõber, Siim, Veldre, Gudrun, Viigimaa, Margus, Levinsson, Anna, Rosengren, Annika, Thelle, Dag S., Hastie, Claire E., Hedner, Thomas, Lee, Wai K., Melander, Olle, Wahlstrand, Björn, Hardy, Rebecca, Wong, Andrew, Cooper, Jackie A., Palmen, Jutta, Chen, Li, Stewart, Alexandre F. R., Wells, George A., **Westra, Harm-Jan**, Wolfs, Marcel G. M., Clarke, Robert, Franzosi, Maria Grazia, Goel, Anuj, Hamsten, Anders, Lathrop, Mark, Peden, John F., Seedorf, Udo, Watkins, Hugh, Ouwehand, Willem H., Sambrook, Jennifer, Stephens, Jonathan, Casas, Juan-Pablo, Drenos, Fotios, Holmes, Michael V, Kivimaki, Mika, Shah, Sonia, Shah, Tina, Talmud, Philippa J., Whittaker, John, Wallace, Chris, Delles, Christian, Laan, Maris, Kuh, Diana, Humphries, Steve E., Nyberg, Fredrik, Cusi, Daniele, Roberts, Robert, Newton-Cheh, Christopher, Franke, Lude, Stanton, Alice V, Dominiczak, Anna F., Farrall, Martin, Hingorani, Aroon D., Samani, Nilesh J., Caulfield, Mark J. & Munroe, Patricia B. Blood pressure loci identified with a gene-centric array. *Am. J. Hum. Genet.* 89, 688–700 (2011).

4

Sampietro, M. Lourdes, Trompet, Stella, Verschuren, Jeffrey J. W., Talens, Rudolf P., Deelen, Joris, Heijmans, Bastiaan T., de Winter, Robbert J., Tio, Rene A., Doevendans, Pieter A. F. M., Ganesh, Santhi K., Nabel, Elizabeth G., **Westra, Harm-Jan**, Franke, Lude, van den Akker, Erik B., Westendorp, Rudi G. J., Zwinderman, Aeilko H., Kastrati, Adnan, Koch, Werner, Slagboom, P. Eline, de Knijff, Peter & Jukema, J. Wouter. A genome-wide association study identifies a region at chromosome 12 as a potential susceptibility locus for restenosis after percutaneous coronary intervention. *Hum. Mol. Genet.* 20, 4748–57 (2011).

3

Anderson, Carl A., Boucher, Gabrielle, Lees, Charlie W., Franke, Andre, D'Amato, Mauro, Taylor, Kent D., Lee, James C., Goyette, Philippe, Imielinski, Marcin, Latiano, Anna, Lagacé, Caroline, Scott, Regan, Amininejad, Leila, Bumpstead, Suzannah, Baidoo, Leonard, Baldassano, Robert N., Barclay, Murray, Bayless, Theodore M., Brand, Stephan, Büning, Carsten, Colombel, Jean-Frédéric, Denson, Lee A., De Vos, Martine, Dubinsky, Marla, Edwards, Cathryn, Ellinghaus, David, Fehrmann, Rudolf S. N., Floyd, James A. B., Florin, Timothy, Franchimont, Denis, Franke, Lude, Georges, Michel, Glas, Jürgen, Glazer, Nicole L., Guthery, Stephen L., Haritunians, Talin, Hayward, Nicholas K., Hugot, Jean-Pierre, Jobin, Gilles, Laukens, Debby, Lawrance, Ian, Lémann, Marc, Levine, Arie, Libioulle, Cecile, Louis, Edouard, McGovern, Dermot P., Milla, Monica, Montgomery, Grant W., Morley, Katherine I., Mowat, Craig, Ng, Aylwin, Newman, William, Ophoff, Roel A., Papi, Laura, Palmieri, Orazio, Peyrin-Biroulet, Laurent, Panés, Julián, Phillips, Anne, Prescott, Natalie J., Proctor, Deborah D., Roberts, Rebecca, Russell, Richard, Rutgeerts, Paul, Sanderson, Jeremy, Sans, Miquel, Schumm, Philip, Seibold, Frank, Sharma, Yashoda, Simms, Lisa A., Seielstad, Mark, Steinhart, A. Hillary, Targan, Stephan R., van den Berg, Leonard H., Vatn, Morten, Verspaget, Hein, Walters, Thomas, Wijmenga, Cisca, Wilson, David C., **Westra, Harm-Jan**, Xavier, Ramnik J., Zhao, Zhen Z., Ponsioen, Cyriel Y., Andersen, Vibeke, Torkvist, Leif, Gazouli, Maria, Anagnou, Nicholas P., Karlsen, Tom H., Kupcinskas, Limas, Sventoraityte, Jurgita, Mansfield, John C., Kugathasan, Subra, Silverberg, Mark S., Halfvarson, Jonas, Rotter, Jerome I., Mathew, Christopher G., Griffiths, Anne M., Gearry, Richard, Ahmad, Tariq, Brant, Steven R., Chamaillard, Mathias, Satsangi, Jack, Cho, Judy H., Schreiber, Stefan, Daly, Mark J., Barrett, Jeffrey C., Parkes, Miles, Annese, Vito, Hakonarson, Hakon, Radford-Smith, Graham, Duerr, Richard H., Vermeire, Séverine, Weersma, Rinse K. & Rioux, John D. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* 43, 246–252 (2011).

2

Hrdlickova, B., **Westra, H. J.**, Franke, L. & Wijmenga, C. Celiac disease: moving from genetic associations to causal variants. Clin. Genet. 80, 203–313 (2011).

1

Szperl, A. M., Ricaño-Ponce, I., Li, J. K., Deelen, P., Kanterakis, A., Plagnol, V., van Dijk, F., **Westra, H. J.**, Trynka, G., Mulder, C. J., Swertz, M., Wijmenga, Cisca & Zheng, H. C. H. Exome sequencing in a family segregating for celiac disease. *Clin. Genet.* 80, 138–47 (2011).

Dankwoord

Onderzoek doe je nooit in je eentje. Er zijn altijd mensen die meer weten dan jij en er zijn altijd mensen die dingen beter doen dan jij. Daarnaast zijn er mensen die je afleiden van de dagelijkse beslommeringen en voor het vieren van feest. Al deze mensen zijn keihard nodig gedurende de vier jaar dat je PhD-student bent en zij verdienen daarom een bedankje.

Beste Lude, baas, ik ben je dankbaar dat ik je eerste PhD student heb mogen zijn. We hebben daardoor de afgelopen 4 jaar veel van elkaar mogen leren. Ik zou hele epistels kunnen wijden aan de pieken en dalen van de afgelopen vier jaren; bijvoorbeeld over de hectiek en euforie rond het submitten van papers vlak voor de kerstvakantie. Helaas heb ik daar echter geen ruimte voor. Desondanks hoop ik dat je de komende jaren je 'eigen ding' blijft doen, zodat er nog vele mooie papers mogen volgen!

Beste Cisca, ik heb je no-nonsense aanpak altijd erg gewaardeerd. Dat komt mogelijk door onze gedeelde Friese achtergrond, al moet daarbij gezegd worden dat jij oorspronkelijk uit de Wâlden komt en ik uit de klei en dat ik je daarom soms niet kon verstaan. Desondanks heb je me altijd gesteund en op een bewonderenswaardig kalme manier kunnen adviseren als ik twijfelde over stappen voor de toekomst, waarvoor ik je erg dankbaar ben.

Daarnaast wil ik graag mijn beoordelingscommissie, Prof. dr. G. de Haan, Prof. dr. L.H. van den Berg en Prof. dr. R.K. Weersma bedanken voor het uitvoerig bestuderen en beoordelen van dit proefschrift.

The best thing about being a PhD-student is that awesome and supportive people surround you. To each and every one of you: thanks for the countless occasions of beer-o-clocks, barbeques and parties that we enjoyed together. Juha, it was a great pleasure to explore and discuss the wonders of the mind and the world around us. Rodgi, thanks for your sharp sense of humor and your awesome music playing skills when recording the Rat Blues. Isis, you are one of the most kind persons in the world that I know and you can cook like crazy. Cleo: sorry that I still can't tell the difference between Limburg and Brabant, but do continue to sing, because it rocks. I especially want to thank Javier: our discussions were rarely scientific, but did include hypothetical letters to the pope, amongst other things. Also, I still think the AssKicker 2000 extravaganza XXL a-gogo is a very good idea (although the mini variant also still sounds appealing). To all of you, I can't wait for the next zombiemovie-marathon.

Danny, de laatste paar maanden dat ik PhD was waren mede door de door jou gedeelde smart een feest, dat we hebben gevierd met het 'weghuffen' van het celtype paper. Ritsert, bedankt dat je altijd bereid was naar mijn papers te kijken en dat je me hebt geleerd kritisch naar de tekst te kijken. Beste Gerard, veel mensen blijven onwetend over het interpreteren van de statistiek en daarom zijn mensen zoals jij hard nodig! Rudolf, ik heb altijd erg genoten van onze discussies over statistiek en Lowlands, maar ik wil je ook erg bedanken voor het carrière advies dat je me her en der gegeven

hebt. Dasha, thanks for the great collaboration on the RNA-seq eQTL analysis. I have learnt a lot about Russia and RNA-seq from you. Dear Vinod, I very much enjoyed our collaboration on the lincRNA eQTL analysis! I hope to read many more eQTL papers from both of you!

Dear Marc-Jan, Marijke, and Jing: although we didn't have airconditioning in our office, and the place was most often very warm, we did have a lot of fun and nice conversations. Dear Patrick, thank you for teaching me a couple of very nifty programming tricks.

Beste Matthieu en Astrid: jullie doen eigenlijk het belangrijkste werk in het lab. Ik ben daarom ook heel blij dat ik nog een blauwe maandag met jullie aan de zuurkast heb mogen staan.

Dear Jackie, everytime I sent you an e-mail I was afraid that my English would hurt your eyes. However, you showed the astonishing ability to stay calm and point out that instead of being greatful, I should be grateful about what you do. So here I am, trying another time to spell it correctly: thank you very much. This thesis would be a Denglish mess without your help. Kate McIntyre: thanks for editing chapter 4 of this thesis.

Beste Helene, Joke, Mentje en Bote, er valt zoveel te regelen in dit vakgebied en jullie zitten daar altijd bovenop. Dank daarvoor.

To all the other people from our group: thank you so much for your support and everlasting kindness.

To all the co-authors on the papers that are presented in this thesis, thank you very much for your collaboration. Out of all these people, my thanks especially go to Tonu Esko, Marjolein Peters, Claudia Schurmann and Katharina Schramm. Tonu: thank you for your endless hospitality during my stay in Boston and your quick response to whatever I sent you. You are truly the good Samaritan of GWAS. Marjolein: thank you very much for your patience, testing the software and writing the accompanying user manual. Claudia and Kathy: thanks a lot for driving me around in Iceland: I never knew you could take a Ford Fiesta off-road!

Dear Soumya and fellow cave dwellers. I am very happy that you have allowed me in your midst and I am sure we will have a great time!

Naast de mensen die je in de academische setting omgeven zijn er ook mensen die er altijd voor je zijn. Lieve Mayke dankjewel voor al je steun, liefde, en elke keer dat je me hebt gekalmeerd als ik weer eens gestrest was om dit proefschrift. Je bent awesome. Beste Max, ik wil niet weten hoe mijn PhD tijd was verlopen zonder de willekeurige park-hang sessies of de bezoekjes aan de Pintelier om de stoom af te blazen. Beste Nikky en lieve Mayra, bedankt voor alle knuffels, lanparty's en (Duitse) biertjes. Dat ghetto-metal-screamo-tec album komt er zeker nog een keer! Beste M. Zuurkool, G. Minstreel en B. Magic, bedankt voor jullie (muzikale) steun. Glijdend van podium naar podium heb ik daarbij genoten van onze band dynamiek. Beste Marcel en Arthur, zonder jullie hadden frituur, mayonaise, curry, uitjes en het neonkruis

nooit zo'n bijzondere betekenis voor me gehad. De week in Appelscha, waar we als ware acid-troubadours de buren lastig vielen door keiharde technobeats, frikadellen, kroketten en mislukte BBQ te droppen, was echt onvergetelijk.

Beste neef Marcel, ontzettend bedankt voor het ontwerpen van mijn proefschrift. We hebben meer gemeen dan je op het eerste gezicht zou denken (naast het typische Westra-hoofd).

Lieve heit en mem, jullie steun en vertrouwen in mij is onbeperkt. Ondanks dat mijn vakgebied soms een ver-van-jullie bed show is, tonen jullie altijd oprechte interesse. Jullie zijn en blijven een groot voorbeeld voor mij. Beste Peter, Tessa, Marijke en in het bijzonder Ytzen, jullie voorspelling dat ik een postbode zou worden is niet correct gebleken, maar de voorspelling dat ik waarschijnlijk een stoffige wetenschapper zou worden met warrig grijs haar en een baard is nu een stapje dichterbij. Ik ben trots op jullie allemaal en blij dat ik jullie als broers en zussen mag hebben.

Harm-Jan Westra
Juli 2014

Colophon

Harm-Jan Westra
Interpreting disease genetics using functional genomics
PhD thesis – Department of Genetics
University of Groningen, University Medical Center Groningen

Layout & Design (with the exception of the figures)
Typysk, Marcel Westra

Printing
Rekladruk, Giekerk

Printing of this thesis was financially supported by:
University of Groningen, University Medical Center Groningen,
the Groningen University Institute for Drug Exploration (GUIDE)
and the Graduate School for Medical Sciences (GSMS), Groningen.

ISBN 978-90-367-7206-8 (print)
ISBN 978-90-367-7205-1 (ebook)